2ª edição

Descobrindo a ESTATÍSTICA usando o SPSS

ANDY FIELD

bookman®

artmed®

ANDY FIELD

Senior Lecturer in Psychology at The University of Sussex, UK

2ª edição

Descobrindo a ESTATÍSTICA usando o SPSS

Tradução, consultoria e supervisão desta edição:

Lorí Viali

Professor Titular da Famat/PUCRS
Professor Adjunto do IM/UFRGS

Versão impressa
desta obra: 2009

2009

Material com direitos autorais

Obra originalmente publicada sob o título *Discovering Statistics with SPSS 2nd Edition*
ISBN 0-71619-4452-4

© English language edition published by Sage Publications of London, Thousand Oaks and New Delhi,
© Andy Field, 2005.
© Portuguese language translation by Artmed Editora 2009.

Capa: *Paola Manica*

Preparação de original: *Maria Francisca de Oliveira Vargas*

Leitura final: *Lara Frichenbruder Kengeriski e Amanda Munari*

Supervisão editorial: *Mônica Ballejo Canto*

Editoração eletrônica: *Techbooks*

Reservados todos os direitos de publicação, em língua portuguesa, à
ARTMED® EDITORA S.A.
Av. Jerônimo de Ornelas, 670 - Santana
90040-340 Porto Alegre RS
Fone (51) 3027-7000 Fax (51) 3027-7070

É proibida a duplicação ou reprodução deste volume, no todo ou em parte, sob quaisquer formas ou por quaisquer meios (eletrônico, mecânico, gravação, fotocópia, distribuição na Web e outros), sem permissão expressa da Editora.

SÃO PAULO
Av. Angélica, 1091 - Higienópolis
01227-100 São Paulo SP
Fone (11) 3665-1100 Fax (11) 3667-1333

SAC 0800 703-3444

IMPRESSO NO BRASIL
PRINTED IN BRAZIL

Material com direitos autorais

Como a primeira edição, este livro é dedicado ao meu irmão Paul e ao meu gato Fuzzy (que é o mesmo gato da última edição, cujo apelido é "Beana", mas que agora é novamente chamado do nome que eu originalmente dei a ele!), porque um deles é uma fonte constante de inspiração intelectual e o outro me acorda de manhã sentando em mim e me lambendo o rosto até que eu lhe dê comida de gato – as manhãs serão muito melhores quando meu irmão superar seu amor por comida de gato no café da manhã.☺

Material com direitos autorais

AGRADECIMENTOS

Obrigado à SPSS pela permissão do uso de suas imagens. A SPSS pode ser contatada na Avenida North Michigan,444, em Chicago, Illinois 60611(Estados Unidos) ou no primeiro andar da St. Andrew's House, West Street, Working, GU21 IEB (UK). Para mais informações, acessem o site (htpp://www.spss.com). Obrigado, também, a Alex Buchner, por disponibilizar seu programa G*Power.

A primeira edição não teria acontecido se não fosse por Dan Wright, que além de ter fé incondicional em um recém-pós-graduado, também leu e revisou muitos capítulos. Esta nova edição recebeu ajuda de muitas pessoas que enviaram *e-mails* oferecendo dicas referentes à primeira edição: agradeço muito a todos que gastaram seu tempo para apontar erros, sugerir melhorias e dar retornos maravilhosos. Assim, meus sinceros agradecimentos vão para Peter de Heus, Tilly Houtmans, Don Hunt, Paul Tinsley, Keith Tolfrey, Jaap Dronkers e Nick Smith (para nomear apenas alguns!). A Sage também organizou várias revisões anônimas que nos deram ótimas sugestões, assim, sou agradecido a essas pessoas, quem quer que elas sejam (Tony Cassidy, que admitiu ser uma delas, recebe meu agradecimento particular!) Jeremy Miles evitou que eu me tornasse um completo idiota (no livro – é uma pena que seus poderes não se estendam para o cotidiano) apontando muitos erros; ele também foi uma pessoa muito agradável de conhecer nestes últimos anos (menos quando ele diz que aquelas seções de rascunhos do meu livro são, e eu cito com suas próprias palavras, "besteira"!). Gareth Williams e Lynne Slocombe ofereceram inspiradas sugestões pedagógicas; entretanto, eu nunca vou perdoá-los por terem sugerido um glossário de termos-chave. Laura também testou a primeira edição do livro como uma tutora na graduação e, mais tarde, na pós-graduação do meu curso; ela também tomou tempo do seu curso de pós-graduação para comentar capítulos desta nova edição e sugerir muitas melhorias. Agradecimentos especiais por esta edição vão para David Hitchin, o herói desconhecido da estatística na University of de Sussex. Ele despendeu muito do seu tempo oferecendo comentários sobre a primeira edição, e sobre alguns rascunhos de capítulos desta nova edição, me ensinando inúmeras coisas sobre estatística por *e-mail*. Todas essas pessoas leram a primeira edição ou grande parte dela (senão toda), e apontaram erros, sugeriram aprimoramentos ou, simplesmente, sugeriram coisas

Material com direitos autorais

que eu antes desconhecia. Não sei o que isso sugere sobre seus estados mentais, mas sou eternamente agradecido por seu bom coração: que eles vivam muito e que seus conjuntos de dados sejam normais.

Com o risco de me prolongar tanto como a cerimônia do Oscar, também estou muito agradecido às seguintes pessoas por escreverem boas críticas à primeira edição, tanto na Amazon.com ou na Amazon.co.uk: George H. Marshall (Escócia), Andrew (Itália), Dr. Simon Marshall (University of Loughborough, Grã Bretanha), um leitor da Royal Holloway University of London, Dr. Keith Tolfrey (Manchester Metropolitan University, Grã Bretanha), James MacCabe (Institute of Psichiatry, Londres, Grã Bretanha), Geoff Bird (University College, Londres, Grã Bretanha), abare@telcel (Grã Bretanha), Jeremy Miles (embora levemente parcial, porque me conhece), Ken Kolosh (IL, EUA), Mehmet Yusuf Yahyagil (Turquia), Sandra Casillas (CA, EUA), um leitor de Missoula (MT, EUA), Paulo Ferreira Leite (Brasil), um leitor de Fair Lawn (NJ, EUA), lts1102 de Cresco (PA, EUA), Shaun Galloway (Hungria), smcol dos EUA, um leitor de Bowling Green (KY, EUA), Denis E. Hommrich (KY, EUA), Hsi Lung Wu (Taiwan), Iwan Wahyu (Singapura) e Mark Gray. Agradeço por vocês escreverem todas essas críticas. Se vocês algum dia forem a Brighton, lhes pago uma cerveja!

O pessoal da Sage é, sem dúvida, o mais beberrão que eu já conheci. Desde a primeira edição, tenho tido a felicidade de trabalhar com Michael Carmichael, que, apesar de suas falhas no campo de futebol (!), participou de memoráveis noitadas comigo. Em seu tempo livre, ele é também um ótimo editor e uma pessoa maravilhosa que merece uma medalha por me aturar. Mas, para ser honesto, eu também mereço uma medalha por aguentar sua obsessão por glossários.

Se você tem um dos meus livros, deve saber que eu sempre escrevo ouvindo música. Devo a minha sanidade, por terem grandes músicas: Fugazi, Beck, Busta Rhymes, Abba, The Cardiacs, Mercury Ver, Ben e Jason, Plug, Roni Size, Supergrass, Massive Attack, Elvis Costello, The Smashing Pumpkins, Radiohead, Placebo, Money Mark, Love, Hefner, Nick Cave, DJ Shadow, Elliott Smith, Muse, Arvo Pärt, AC/DC e Quasi. Para essa segunda edição, a minha lista de músicas é um indicativo da angústia mental na qual afundei. Os tons encantadores das seguintes músicas me envolveram: Emperor, Cradle of Fifth, Frank Black and the Catholics, Blondie, Fugazi, Radiohead, Peter Gabriel, Genesis (a era de Peter Gabriel – tem um pouco da fase de rock progressivo por volta do Capítulo 5), Metallica, The White Stripes, Sevara Nazarkhan, Nusrat Fateh Ali Khan, Killing Joke, The Beyond, Jane's Addiction, Nevermore, The French and Hefner, Iron Maiden, The Mars Volta, Morrisey, Slipknot, Granddaddy, Mark Lanegan, PJ Harvey.

Escrever um livro exige horas solitárias (geralmente tarde da noite) de digitação. Sem a ajuda de meus amigos que me tiravam da minha sala pouco iluminada de tempos em tempos, seria mais abobado do que eu já sou. Em especial, sou grato a Graham e Benie, Martin, Doug, Paul, Darren, Helen Liddle e Mark por me lembrarem que existe vida além da estatística, e à Leonora, por achar que psicólogos que gostam de estatística dão bons maridos.☺

Material com direitos autorais

PREFÁCIO

Policial, prenda este homem,
ele fala em números
Ele zune como uma geladeira, ele
é como um rádio fora do ar.

(Radiohead, 1997)

INTRODUÇÃO

Os estudantes de ciências sociais sempre desprezaram a Estatística. Uma das razões é que muitos não entendem matemática, o que torna a compreensão de equações estatísticas complexas muito difícil. A maior vantagem em ter estudado estatística no início dos anos de 1990 (como eu estudei) comparado com os anos de 1960 foi o desenvolvimento de softwares para fazer todo o trabalho. A vantagem em aprender estatística hoje, comparado a dez anos atrás, é que esses pacotes agora são mais fáceis de usar graças ao Windows®/MacOS®. O SPSS é, em minha opinião, o melhor dos pacotes comerciais disponíveis, utilizado em muitas universidades. Então, o que me levou a escrever um livro sobre estatística e SPSS?

Já existem muitos bons livros que descrevem a teoria estatística. Howell (2002), Stevens (1992) e Wright (2002) escreveram livros maravilhosos e claros, mas utilizam exemplos de computador somente como um adendo à teoria. Da mesma maneira, existem muitos bons livros sobre SPSS (Kinnear e Gray, 2000; Foster, 2001), mas eles se concentram em "fazer os testes". Usar SPSS sem conhecimento algum de estatística pode ser perigoso (é somente uma ferramenta, não um recurso divino de sabedoria). Portanto, quero usar o SPSS como uma ferramenta para ensinar conceitos estatísticos. Agindo dessa maneira, espero que o leitor tenha um entendimento melhor, tanto da teoria quanto da prática.

Primeiro, quero responder o tipo de pergunta que eu fazia quando estudava estatística e usava o SPSS na faculdade (por exemplo, "Será que consigo entender como funciona esse teste estatístico sem saber muito de matemática?", "O que este botão faz?", "Mas que diabos significa esta saída?"). O SPSS tem um complexo conjunto de opções para cada teste, muitos dos quais são ensinados de modo superficial por livros e professores. Espero ser capaz de explicar o que essas opções realmente fazem e por que você deve usá-las. Quero ir além de fornecer receitas. Muitos livros dizem ao leitor o que fazer ("aperte esta tecla", "faça isso", "faça aquilo", etc.), e isso cria uma impressão de que a estatística e o SPSS são infle-

Material com direitos autorais

xíveis. O SPSS tem muitas opções projetadas que lhe permitem modificar um teste já elaborado para seus fins específicos. Portanto, embora eu faça recomendações, espero fornecer aos leitores fundamentos teóricos suficientes para torná-los capazes de tomar suas próprias decisões sobre quais opções são mais apropriadas para as análises que eles querem fazer.

Meu segundo objetivo foi escrever um livro que pudesse ser lido por vários níveis (veja a próxima seção para mais detalhes). Existem capítulos para alunos do primeiro ano da faculdade (1, 2, 3, 4, 7 e 13), e capítulos para estudantes do segundo ano (5, 8, 9, 10, 11 e 12) e capítulos de tópicos mais avançados que alunos pós-graduados poderão usar (6, 14, 15 e 16). Todos esses capítulos devem ser acessíveis a qualquer um e espero conseguir isso indicando o nível adequado de cada seção (veja a próxima).

O QUE HÁ DE NOVO?

Eu era muito ruim em matemática. Aos 13 anos, era praticamente o último da minha turma. No entanto, 12 anos depois escrevi um livro-texto sobre estatística (e 17 anos depois terminei a segunda edição). A diferença entre o menino de 13 anos que não passou no seu exame e o de 15 anos que foi muito bem foi um bom professor: meu irmão, Paul. De fato, devo minha vida acadêmica à habilidade de Paul em fazer o que meus professores não fizeram: ensinar as matérias de uma maneira envolvente. Ainda hoje ele aparece, quando necessário, para me ensinar coisas (um curso intensivo de programação de computador, alguns Natais atrás, me vem à lembrança). De qualquer modo, a razão de ele ser um bom professor é que ele é capaz de tornar as coisas interessantes e relevantes para mim. Infelizmente, ele parece ter os genes de "bom professor" da família (e ele nem mesmo trabalha como professor, que desperdício!), mas eu tenho tentado usar a sua abordagem nas minhas aulas e livros. Algo que aprendi foi que as pessoas apreciam um toque de humanidade, assim, na primeira edição, tentei colocar muito da minha personalidade (ou seria "falta de personalidade?"). Então, havia alguns exemplos despretensiosos (alguns disseram "indecentes", mas prefiro "despretensiosos") e uma grande dose de humor – o que pode ter sido uma coisa ruim porque eu sou um professor e, portanto, não tenho senso de humor!

Quando escrevi a primeira edição, meu objetivo era escrever um tipo de livro de estatística que eu gostaria de ler. Muito egoísta, eu sei, mas achei que se eu tivesse um livro de referência com exemplos que me divertissem, a vida se tornaria mais simples quando eu necessitasse de um livro para consulta. Não achei que alguém fosse comprar a obra (bem, fora minha mãe e meu pai) e eu antecipei desculpas para comentários do tipo "todo o capítulo X está completamente errado e você é um idiota", ou "pense em quantas árvores morreram para este lixo ser escrito, você deveria se envergonhar". Na verdade, nem a editora pensou que o livro iria vender (eles revelaram isso posteriormente, devo acrescentar). Eu realmente não esperava receber muitos *e-mails* extremamente gentis de pessoas que gostaram do livro. Até hoje ainda fico admirado que alguém leia o livro e gaste tempo para me escrever um *e-mail* gentil (e devo acrescentar, com risco de parecer sentimental, que saber que o livro ajuda as pessoas faz com que o sorriso na minha face nunca desapareça). Mas por que estou contando tudo isso? Bem, vendo que pelo menos algumas pessoas apreciaram o estilo da primeira edição, tomei isto como uma luz verde para incluir mais exemplos estúpidos, mais bobagens e mais mau gosto. Resumindo, muito mais sexo, drogas e *rock 'n' roll*. Para todos aqueles que odeiam isso, desculpa, mas isso me diverte.

Além de acrescentar mais bobagens, fui forçado, com relutância, a expandir o conteúdo acadêmico. Muitos dos acréscimos são resultados de alguém (geralmente muitas pessoas) que me enviaram *e-mails* perguntando como fazer algo. Assim, teoricamente, esta edição deveria responder as perguntas que me foram feitas nos últimos quatro anos! As mudanças gerais em cada capítulo são:

Material com direitos autorais

- **Exemplos**: Cada capítulo agora tem muitos exemplos no final para você trabalhar, incluindo as respostas.
- **Mais dados**: Por causa dos exemplos, você tem mais conjuntos de dados!
- **Relatando suas análises:** Existem seções sobre como relatar as análises na maioria dos capítulos. Essas seções foram baseadas no manual de estilo de publicação da APA (American Psychological Association, 2001).
- **Glossário:** Termos-chave estão agora identificados no texto e existe um glossário de todos eles.
- **Tamanho de efeito:** A maioria dos tópicos agora tem uma discussão sobre como calcular o tamanho de efeito.
- **Resumos:** Há muita informação no texto e meu estilo é divagador (prefiro "conversador"), assim, existem agora quadros com resumo dos pontos-chave (veja a próxima seção sobre como usar este livro).
- **Quadros:** Inúmeros quadros que ressaltam temas interessantes.
- **Sintaxe:** Existe agora discussão sobre o uso da sintaxe do SPSS ao longo do livro.
- **SPSS 13:** Embora essas fossem as menores mudanças, porque o SPSS 13 não é muito diferente da versão 9, atualizei tudo com a última versão do SPSS.

As mudanças específicas de cada capítulo são:

- **Capítulo 1 (Conceitos básicos):** Foi expandido para incluir mais fundamentos em teoria da estatística (por exemplo, intervalos de confiança, testes de significância, poder, tamanho de efeito).
- **Capítulo 2 (SPSS):** Estava condensado dentro do Capítulo 1, mas foi atualizado para lidar com o SPSS 13 e agora inclui discussão da assustadora janela de sintaxe.
- **Capítulo 3 (Explorando dados):** Tem alguma semelhança com o antigo Capítulo 2 da primeira edição. Ele agora aborda apresentação de dados, transformação de dados, verificação de modelos de dados e muito mais.
- **Capítulo 4 (Correlação):** Estava no Capítulo 3 e, na verdade, não mudou muito.
- **Capítulo 5 (Regressão):** Estava no Capítulo 4 e agora tem um material mais avançado, como uma seção em previsores categóricos, bem como uma discussão de autovalores. Também expandi as seções originais de forma mais técnica (por exemplo, uma discussão mais aprofundada do tamanho de efeito).
- **Capítulo 6 (Regressão Logística):** O antigo Capítulo 5 foi aumentado levemente de várias formas (a parte teórica principalmente), mas também reorganizei alguns conteúdos de modo que a teoria e as dicas do SPSS estão mais claramente separadas.
- **Capítulo 7 (Testes t):** Não foi alterado em relação ao antigo Capítulo 6 exceto algumas informações das quais não gostava que foram reescritas.
- **Capítulo 8 (MLG 1):** É basicamente o mesmo antigo Capítulo 7, exceto que agora cobre as versões de Welch e Brown-Forsythe da F quando a homogeneidade da variância não pode ser assumida.
- **Capítulo 9 (MLG 2):** Aborda a análise de covariância (ANCOVA), que antes dividia um capítulo com a ANOVA Fatorial, mas agora ganhou um capítulo próprio. Coloquei algum material extra sobre contrastes planejados e corrigi alguns erros embaraçosos da primeira edição.☺
- **Capítulo 10 (MLG 3):** Cobre ANOVA Fatorial (que dividia um capítulo com a ANCOVA). Existem seções inteiramente novas sobre a teoria da ANOVA independente de dois fatores, como realizar contrastes (com sintaxe), como realizar análise de efeitos simples, interpretação de diagramas de interações e como podemos conceitualizar a ANOVA de dois fatores como um MLG (Modelo Linear Generalizado).
- **Capítulo 11 (MLG 4):** Era o Capítulo 9 sobre a ANOVA de medidas repetidas. Adicionei uma seção sobre a teoria da ANOVA de medidas repetidas e sobre a

Material com direitos autorais

condução de análises de efeitos simples na ANOVA de medidas repetidas com dois fatores.

- **Capítulo 12 (MLG 5):** Aborda a ANOVA mista (estava junto com ANOVA de medidas repetidas no antigo Capítulo 9). Existe um exemplo completamente novo e, com o novo capítulo, expandi um pouco as seções de interpretação.
- **Capítulo 13 (Estatística não-paramétrica):** É um capítulo completamente novo (embora pegue emprestadas algumas coisas do Capítulo 2). Acrescentei seções de teoria para os testes de Mann-Whitney e Wilcoxon; também acrescentei seções completamente novas que cobrem a ANOVA de Friedman, o teste de Jonckheere e o teste de Kruskal-Wallis.
- **Capítulo 14: (MANOVA):** Não mudou muito do Capítulo 10.
- **Capítulo 15 (Análise de fatores):** Reorganizei o material para que o que estava escondido na seção de interpretação ficasse agora claramente colocado na seção de teoria. É uma seção completamente nova sobre análise de confiabilidade.
- **Capítulo 16 (Dados categóricos):** É basicamente um capítulo novo (embora o exemplo do qui-quadrado seja antigo). Incluí uma seção de teoria para o teste qui-quadrado, razão de verossimilhança e correção de Yates, discussão sobre razão de chances (*odds ratios*) e análise Loglinear. Nenhuma dessas seções estava na antiga edição!

ADEUS

Escrevi no início da primeira edição que "este livro é o resultado de dois anos (com uma ou duas semanas para escrever a minha tese de doutorado) tentando atingir esses objetivos. Ele não é perfeito, e eu gostaria de ter um retorno (bom ou ruim) das pessoas que realmente contam: vocês, os leitores". Esse sentimento ainda se aplica, exceto que agora é o resultado de dois anos e meio de trabalho. Ao longo desses últimos quatro anos, fiquei muito ligado a este livro: ele começou com um trabalho de amor e ainda é. Com o inesperado sucesso da primeira edição, e, tendo feito muitas atualizações para a segunda edição (trezentas páginas comparadas com as cinquenta páginas que a editora queria que eu escrevesse!), estou muito preocupado que eu tenha mudado tudo! Portanto, embora eu não saiba se o livro terá uma terceira edição, mesmo assim, gostaria de ter o retorno das pessoas que realmente contam: vocês, os leitores.

Andy

E-mail: discoveringstatistics@sussex.ac.uk
Web: http/www.sussex.ac.uk./Users/andyf/

Material com direitos autorais

COMO USAR ESTE LIVRO

Quando a editora me pediu para escrever uma seção sobre "como usar este livro" foi tentador escrever "compre um frasco de creme antirrugas (o que você irá precisar para se defender dos efeitos do envelhecimento enquanto ler o livro), procure uma cadeira confortável, sente, dobre a capa, comece a ler e só pare quando chegar à última página". Entretanto, acho que eles gostariam de algo mais útil.☺

QUE CONHECIMENTO ANTERIOR VOCÊ PRECISA TER?

Baseei-me na suposição de que você não sabe nada sobre estatística, mas tem algum conhecimento básico sobre computador (eu não vou dizer como ligá-lo, por exemplo) e matemática (embora eu tenha incluído uma revisão rápida sobre alguns conceitos básicos, assim, eu não suponho nenhum conhecimento prévio).

OS CAPÍTULOS FICAM MAIS DIFÍCEIS AO LONGO DO LIVRO?

De certa forma ficam (o Capítulo 14 sobre MANOVA é mais difícil do que o Capítulo 1), mas em outros casos, não (o Capítulo 13, sobre estatística não-paramétrica é menos complexo que o Capítulo 12, e o Capítulo 7, sobre o teste-t, é menos complexo do que Capítulo 6 sobre regressão logística).

Por que eu fiz isso? Ordenei os capítulos para ter coerência estatística (para mim, pelo menos). Muitos livros ensinam vários testes isoladamente e não apresentam a idéia global; isso, eu acho, cria um mistério desnecessário. A maioria dos testes neste livro é a mesma apresentada de uma maneira ligeiramente diferente. Assim, eu queria que o livro contasse essa história. Para fazer isso, tenho que explicar alguns conceitos, como o de regressão, bem no começo, porque quase tudo está baseado nesse conhecimento!

Entretanto, para ajudá-lo, codifiquei cada seção com um ícone, que representa a dificuldade do conteúdo. Eles não significam que você deva pular seções (veja o Alex Esperto na próxima seção), mas irão informar se a seção está no seu nível ou se irá exigir mais conhecimento de você. Eu baseei esses ícones no meu modo de ensinar, portanto, eles podem não se adequar a todos (especialmente porque sistemas variam em diferentes países!)

① Isto significa "nível um" e eu o equiparei ao primeiro ano da faculdade no Reino

Material com direitos autorais

Unido. Todos deveriam ser capazes de entender essas seções.

② Este é o próximo nível e eu o equiparei ao segundo ano da faculdade no Reino Unido. São tópicos que eu leciono no meu segundo ano e qualquer um com pouco de conhecimento de estatística deveria ser capaz de entender. Entretanto, algumas dessas seções são desafiadoras mesmo para alunos do segundo ano. São seções intermediárias.

③ Este é o "nível 3" e representa tópicos difíceis. Espero que alunos do último ano do Reino Unido e recém-formados sejam capazes de lidar com essas seções.

④ Este é o nível mais alto e representa tópicos muito difíceis. Espero que esses tópicos sejam muito desafiadores para estudantes da graduação e recém-formados, mas os recém-formados, com um razoável conhecimento em métodos de pesquisa, não deverão ter muitas dificuldades.

POR QUE VOCÊ VÊ ROSTOS ESTÚPIDOS EM TODOS OS LUGARES?

Alex Esperto: Alex é um personagem muito importante porque ele aparece quando as coisas ficam difíceis. Ele é um pouco malandro, assim, quando a sua cara aparecer, você saberá que algo complicado irá ser explicado. Quando o conteúdo difícil termina, ele reaparece para informar que é seguro continuar. Mas isso não quer dizer que o resto do material do livro é fácil; ele apenas informa as partes do livro que você pode pular se tem coisas melhores a fazer com sua vida do que ler 800 páginas! Então, quando o Alex Esperto aparecer, você pode *pular a seção inteira* e ainda assim entender o que está acontecendo. O Alex também aparece no final de cada capítulo apresentando algumas tarefas para testar se você é tão esperto quanto ele. A propósito, qualquer semelhança entre Alex e o meu editor é pura coincidência!

Samanta Ferrinho: Samanta odeia estatística. Na verdade, ela acha que tudo isso é uma perda de tempo. Ela simplesmente quer passar no exame e esquecer que teve de aprender distribuição normal. Assim, quando aparece, ela fornece um resumo dos pontos-chave que você precisa saber. A Samanta lhe dará a informação essencial para que você não fique procurando em centenas de páginas o que deseja saber.

Brian Cansado: Brian aparece fazendo perguntas e completamente confuso. Portanto, ele não é muito diferente do autor. À medida que o livro avança, ele se torna cada vez mais desanimado. Tire suas próprias conclusões!

Gato Curioso: Ele também aparece para fazer perguntas (porque ele é curioso). Na verdade, ele só está aqui é porque eu queria um gato no livro… e, de preferência, um que se parecesse com o meu. É claro que os especialistas em educação acham que ele precisa de um papel específico, assim, seu papel é parecer bonitinho e contar péssimas piadas sobre gatos.

PARA QUE SERVEM OS QUADROS?

Há quadros ao longo de todo o livro. Os quadros podem ser ignorados, se você quiser, porque são lembretes incluídos no texto. Eles explicam assuntos complicados ou ilustram pontos interessantes sobre estatística, com os quais você pode impressionar seus amigos.

O QUE ESTÁ NO *SITE**?

O *site* contém muitas coisas interessantes:

- **Arquivo de dados:** Em geral, você irá usá-lo nos exemplos, porque ele contém todos os arquivos de dados dos exemplos

* N. de R. O conteúdo disponibilizado no site www.artmed.com.br está em inglês. Consulte também: www.sagepub.co.uk/field, que contém materiais adicionais para o estudante (conteúdo em inglês).

Material com direitos autorais

do livro (os dados para cada capítulo estão contidos numa pasta separada).

- **Respostas para o Alex Esperto:** Porque não quero matar mais árvores do que o necessário, há centenas de materiais adicionais no *site*. Os documentos mais importantes são as resposta às tarefas no final de cada capítulo. Para cada tarefa, o documento contém a saída do SPSS e uma breve explicação, assim, você pode checar suas respostas.
- **Material adicional:** Para alguns tópicos, escrevi longos documentos. São, principalmente, tópicos avançados ou tópicos para os quais uma explicação detalhada no corpo do livro seria uma distração. Em caso de alguém querer saber mais, há descrições no *site*. Por exemplo, há arquivos sobre contrastes usando a sintaxe, a razão F de Welch, teste de Jonckheere e assim por diante. A maioria pode ignorar esses arquivos, mas eles estão no livro caso alguém esteja interessado.
- **Material do apêndice:** Algum material do apêndice (como os cálculos matemáticos) também foi colocado nos arquivos do *site* para economizar papel.
- **Programas:** Há uma cópia do programa G*Power no *site*.

Boa leitura!

Material com direitos autorais

SÍMBOLOS UTILIZADOS

Operadores matemáticos

Σ	Este símbolo (denominado sigma) significa "some tudo". Assim, se você vir algo como Σx_i isso quer dizer "some todos os valores que coletou".
Π	Este símbolo significa "multiplique tudo". Assim, se aparecer algo como Πx_i isso apenas quer dizer que você deve "multiplicar todos os valores que coletou".
$\sqrt{x}$	Isto quer dizer "tire a raiz de x".

Símbolos gregos

α	É a probabilidade de cometer erro do Tipo I
β	É a probabilidade de cometer erro do Tipo II
β_i	Coeficiente da regressão padronizado
χ^2	Estatística teste qui-quadrado
χ^2_F	Estatística teste ANOVA de Friedman
ε	Normalmente significa "erro"
η^2	Eta ao quadrado (uma medida do tamanho de um efeito)
μ	A média de uma população de valores
ρ	O coeficiente de correlação populacional
σ^2	A variância de um conjunto de dados populacionais
σ	O desvio padrão de um conjunto de dados populacionais
σ_X	O erro padrão da média
τ	Tau de Kendall (coeficiente de correlação não-paramétrico)
ω^2	Ômega ao quadrado (medida do tamanho de efeito)

Símbolos ingleses

b_i	O coeficiente de regressão (não-padronizado)
gl	Graus de liberdade (df = *degrees of freedom*)
e_i	O erro associado à i-ésima pessoa
F	Razão F (estatística teste utilizada na ANOVA)
H	Estatística teste de Kruskal-Wallis
k	O número de níveis de uma variável (isto é, o número de condições de um tratamento) ou o número de previsores num modelo de regressão
ln	Logaritmo natural
MS	O erro médio ao quadrado (média ao quadrado). A variabilidade média nos dados

Material com direitos autorais

N, n, n_i	O tamanho da amostra. N normalmente representa o tamanho de uma amostra qualquer enquanto n representa o tamanho de uma particular amostra.
P	Probabilidade (o valor de uma probabilidade, o valor-p ou a significância de um teste são normalmente representados por p)
r	O coeficiente de correlação de Pearson
r_b, r_{pb}	Coeficiente de correlação bisserial e coeficiente de correlação ponto-bisserial, respectivamente
r_s	Coeficiente de correlação por postos de Spearman
R	O coeficiente de correlação múltiplo
R^2	O coeficiente de determinação (isto é, a proporção da variância dentro de alguns dados explicados pelo modelo)
s	O desvio padrão de uma amostra
s^2	A variância de uma amostra
SS	A soma dos quadrados ou a soma dos erros ao quadrado se for dado o nome completo
SS_A	A soma dos quadrados para a variável A
SS_M	A soma dos quadrados do modelo (isto é, a variabilidade explicada pelo modelo ajustado aos dados)
SS_R	A soma dos quadrados dos resíduos (isto é, a variabilidade que o modelo não pode explicar, o erro no modelo)
SS_T	A soma total dos quadrados (isto é, a total variabilidade dos dados)
t	A estatística teste de Student
T	A estatística para o teste de postos com sinais de Wilcoxon
U	A estatística para o teste de Mann-Whitney
W_S	A estatística para o teste da soma dos postos de Wilcoxon
$\overline{X}$ ou $\bar{x}$	A média de uma amostra
z	Um ponto de dados expresso em unidades de desvio padrão

Revisão matemática

1. **Dois negativos fazem um positivo**: embora na vida dois erros não façam um acerto, na matemática eles fazem! Quando multiplicamos um número negativo por outro negativo, o resultado é um número positivo. Por exemplo, $-2 \times -4 = 8$.
2. **Um número negativo multiplicado por um positivo produz um negativo**: se você multiplica um número positivo por um negativo, o resultado é outro número negativo. Por exemplo, $2 \times -4 = -8$ ou $-2 \times 6 = -12$
3. **PODMAS**: é um acrônimo para a ordem em que as operações matemáticas são executadas. Ele significa Parênteses, Ordem, Divisão, Multiplicação, Adição e Subtração, e essa é a ordem em que você deve executar operações em uma expressão. Muitas dessas operações são autoexplicativas (por exemplo, sempre calcule coisas dentro dos parênteses primeiro), exceto por ordens que de fato se referem a termos com potências como quadrados. Quatro ao quadrado, ou 4^2, é lido como quatro elevado na ordem dois, por isso esses termos são denominados "ordem" em PODMAS (também, se os denominarmos de potência, vamos ter PPODMAS, o que será difícil de pronunciar). Vamos dar uma olhada em um exemplo de PODMAS: qual seria o resultado de $1 + 3 \times 5^2$? A resposta é 76 (não 100, como alguns de vocês podem ter pensado). Não existem parênteses, assim, o primeiro a fazer é lidar com o termo ordem: $5^2 = 25$, assim a expressão se torna $1 + 3 \times 25$. Não existe divisão, portanto, podemos passar para a multiplicação: 3×25, que resulta 75. PODMAS nos diz que agora é hora de lidarmos com a adição: $1 + 75$, que fornece o resultado 76. Se eu tivesse escrito a expressão original como $(1 + 3) \times 5^2$, a resposta seria 100 porque nesse caso devemos lidar primeiro com os parênteses: $(1 + 3) = 4$, assim a expressão torna-se 4×5^2. Agora devemos lidar com o termo ordem e a expressão torna-se: $4 \times 25 = 100$!
4. http://www.easymaths.com é um bom local para uma revisão de conceitos matemáticos básicos.

Material com direitos autorais

COMENTÁRIOS SOBRE A PRIMEIRA EDIÇÃO

DE ESTUDANTES:

O livro está muito bem escrito. Está tão claro que é como se eu visse a estatística, que eu estudei por tanto tempo, através de um par de óculos limpos.
Megan Gray, Universidade de Sussex

Sua combinação da parte técnica com o conceito é simplesmente maravilhosa. E o seu senso de humor, também!
Abigail Levy

Você tem a habilidade de explicar coisas complexas de uma maneira simples e compreensível.
Ruth Mann

Eu não seria capaz de terminar meus estudos tão facilmente se não fosse o seu livro.
Marleen Smits, Universidade de Maastricht

Seu livro me tornou um aluno muito mais feliz.
Lisa Oliver, Universidade da Columbia Britânica

Eu nunca vi conceitos tão bem explicados ou ilustrados tão claramente.
Joel Philip

Eu defendi minha dissertação em psicologia em março e seu livro SALVOU MINHA VIDA!
Noelle Leonard

Eu acho que você vai para história como a primeira pessoa a colocar humor em um livro de estatística!
Carol McSweeney, Universidade Brookes

Você é a mais pura expressão de um bom professor! Embora eu não o conheça ou tenha assistido às suas aulas, sua empatia e seu desejo genuíno de ajudar os estudantes se refletem no livro.
Johannah Sirkka

DE ACADÊMICOS

Ano passado descobri suas páginas na rede (e, desse modo, seu livro) e ele transformou completamente minhas aulas de estatís-

Material com direitos autorais

tica básica para alunos de pós-graduação de "ok" para "ótimas". Como resultado, este ano (e num futuro próximo) usarei o seu livro como texto principal na minha classe.

Michael Marsiske, Universidade da Flórida

Eu recomendei seu livro para um dos meus cursos e estou ansioso para que meus alunos cheguem. Leciono estatística há 25 anos e usei muitos livros. Mas este é o melhor. Você me deu uma nova visão de coisas com as quais tive dificuldade por anos.

Henry Steel, Universidade de Stellenbosch, África do Sul

A maneira como você consegue trazer o leitor ao mundo da estatística avançada sem muito jargão técnico é admirável.

Michael A. Karchmer, Diretor do Instituto de Pesquisa Gallaudet, Universidade de Gallaudet

Muito obrigado por escrever um livro excelente e inspirador sobre um tópico que é quase impossível tornar interessante – você fez um milagre!

Dr. Keith Tolfrey, Professor Sênior, Departamento da Ciência do Esporte e do Exercício, Universidade Metropolitana de Manchester.

Desde que o descobri, recomendei seu livro para os meus colegas e acho que pelo menos metade do departamento tem seu livro na estante agora. Quando alguém não sabe o que fazer com seus dados, simplesmente dizemos: consulte o Andy!

Froukje Dijk, Universidade de Maastricht

Material com direitos autorais

SUMÁRIO

AGRADECIMENTOS vii
PREFÁCIO ix
COMO USAR ESTE LIVRO xiii
SÍMBOLOS UTILIZADOS xvi
COMENTÁRIOS SOBRE A PRIMEIRA EDIÇÃO xviii

1 TUDO O QUE VOCÊ SEMPRE QUIS SABER SOBRE ESTATÍSTICA (BEM, QUASE TUDO) 31
1.1 O que você vai aprender neste capítulo? ① 31
1.2 Construindo modelos estatísticos ① 31
1.3 Populações e amostras ① 33
1.4 Modelos estatísticos simples ① 33
1.4.1 A média, a soma dos quadrados, a variância e o desvio padrão ① 33
1.5 Distribuições de frequências ① 37
1.5.1 Propriedades das distribuições de freqüências ① 37
1.5.2 O desvio padrão e a forma da distribuição ① 38
1.5.3 O que é a distribuição normal padrão? ① 39
1.6 A minha amostra é representativa da população? ① 42
1.6.1 O erro padrão ① 42
1.6.2 Intervalos de confiança ② 44
1.7 Modelos lineares ① 47
1.8 Como descobrir se o seu modelo representa o mundo real? ① 49
1.8.1 Estatísticas teste ① 53
1.8.2 Testes uni e bilaterais ① 54
1.8.3 Erros do tipo I e do tipo II ① 56
1.8.4 Tamanhos de efeito ② 56
1.8.5 Poder estatístico ② 57
1.9 Últimos conselhos 58
1.10 O que descobrimos sobre estatística? ① 59
1.11 Termos-chave que descobrimos 59
1.12 Questões de revisão do Alex Esperto 59
1.13 Leituras complementares 60

2 O AMBIENTE DO SPSS 61
2.1 O que você vai aprender neste capítulo? ① 61
2.2 Versões do SPSS ① 61
2.3 Iniciando ① 62
2.4 O editor de dados ① 62
2.4.1 Entrando com os dados no editor ① 68
2.4.2 Criando uma variável ① 69
*2.4.3 O visualizador de variáveis (*Variable View*)* ① 70
2.4.4 Criando variáveis codificadas ① 72
2.4.5 Tipos de variáveis ① 74
*2.4.6 Valores que faltam (*missing*)* ① 75

Material com direitos autorais

2.4.7 Mudando o formato da coluna ① 76
2.4.8 Teste rápido ① 77
2.5 O visualizador de saídas (*Output Viewer*) ① 77
2.6 A janela de sintaxe ③ 80
2.7 Salvando arquivos ① 81
2.8 Recuperando um arquivo ① 82
2.9 O que descobrimos sobre estatística? ① 83
2.10 Termos-chave que descobrimos 83
2.11 Tarefa do Alex Esperto 83
2.12 Leituras complementares 84

3 EXPLORANDO DADOS 85
3.1 O que você vai aprender neste capítulo? ① 85
3.2 Dados paramétricos ① 85
3.2.1 Hipóteses "retirar" ① 85
3.3 Apresentando dados graficamente ① 87
3.3.1 Passo 1: Apontando os erros óbvios utilizando histogramas ① 87
3.3.2 Passo 2: Estatísticas descritivas e o diagrama de caixa e bigodes ① 92
3.3.3 Passo 3: Corrigindo problemas nos dados ② 99
3.3.4 Passo 4: Transformando dados utilizando o SPSS ② 101
3.4 Explorando grupos de dados ① 107
3.4.1 Executando a análise para todos os dados ① 107
3.4.2 Saída do SPSS para todos os dados ① 108
3.4.3 Executando a análise para grupos diferentes ① 110
3.4.4 Saídas para grupos diferentes ① 110
3.5 Testando se uma distribuição é normal ① 112
3.5.1 Executando o teste de Kolmogorov-Smirnov no SPSS ① 113
3.5.2 Saída do procedimento Explorar (Explore) ① 113
3.6 Testando a homogeneidade da variância ① 116
3.7 Representando médias graficamente ① 119
3.8 O que descobrimos sobre estatística? ① 123
3.9 Termos-chave que descobrimos 124
3.10 Tarefas do Alex Esperto 124
3.11 Leituras complementares 124

4 CORRELAÇÃO 125
4.1 O que você vai aprender neste capítulo? ① 125
4.2 Como medimos relacionamentos? ① 125
4.2.1 Um desvio para o mundo da covariância ① 125
4.2.2 Padronização e o coeficiente de correlação ① 128
4.3 Entrada de dados para realizar uma análise de correlação no SPSS ① 129
4.4 Representando relacionamentos graficamente: o diagrama de dispersão ① 129
4.4.1 Diagrama de dispersão simples ① 130
4.4.2 Diagrama de dispersão tridimensional ① 133
4.4.3 Diagrama de dispersão sobreposto ① 134
4.4.4 Diagrama de dispersão matricial ① 136
4.5 Correlação bivariada ① 137
4.5.1 Coeficiente de correlação de Pearson ① 140
4.5.2 Um alerta sobre interpretação: Causalidade ① 142
4.5.3 Utilizando o R^2 para interpretação ① 143
4.5.4 Coeficiente de correlação de Spearman ① 144
4.5.5 Tau de Kendall (não-paramétrico) ① 145
4.5.6 Correlações bisserial e bisserial por ponto ② 146
4.6 Correlação parcial ② 148
4.6.1 A teoria por trás da correlação parte e parcial ② 148
4.6.2 Correlação parcial utilizando o SPSS ② 149
4.6.3 Correlações semiparciais (ou por parte) ② 153
4.7 Como relatar coeficientes de correlação ① 153
4.8 O que descobrimos sobre estatística? ① 154
4.9 Termos-chave que descobrimos 155

Material com direitos autorais

4.10 Tarefas do Alex Esperto 155
4.11 Leituras complementares 155

5 REGRESSÃO 156
5.1 O que você vai aprender neste capítulo? ① 156
5.2 Uma introdução à regressão ① 156
5.2.1 Algumas informações importantes sobre retas ① 157
5.2.2 O método dos mínimos quadrados ① 158
5.2.3 Avaliando o ajuste: soma dos quadrados, R *e* R^2 ① 159
5.2.4 Interpretando previsões individuais ① 162
5.3 Executando regressão simples no SPSS ① 164
5.4 Interpretando a regressão simples ① 165
5.4.1 Ajuste global do modelo ① 165
5.4.2 Parâmetros do modelo ① 166
5.4.3 Utilizando o modelo ① 167
5.5 Regressão múltipla: o básico ② 168
5.5.1 Um exemplo de um modelo de regressão múltipla ② 168
5.5.2 Soma dos quadrados, R *e* R^2 ② 170
5.5.3 Métodos de regressão ② 170
5.5.3.1 Hierárquico (Entrada em blocos) ② 170
5.5.3.2 Entrada forçada ② 171
5.5.3.3 Métodos por passos (Stepwise) ② 171
5.5.3.4 Escolhendo um método ② 172
5.6 Quão acurado é o meu modelo de regressão? ② 172
5.6.1 Interpretando o modelo de regressão I: diagnósticos ② 172
5.6.1.1 Valores atípicos (*outliers*) e resíduos ② 172
5.6.1.2 Casos influentes ③ 174
5.6.1.3 Um comentário final sobre estatísticas diagnóstico ② 178
5.6.2 Interpretando o modelo de regressão II: generalização ② 178
5.6.2.1 Avaliando hipóteses ② 178
5.6.2.2 Validação cruzada do modelo ③ 180
5.6.2.3 Tamanho da amostra na regressão ③ 181
5.6.2.4 Multicolinearidade ② 182
5.7 Como executar uma regressão múltipla no SPSS ② 184
5.7.1 Principais opções ② 184
5.7.2 Estatísticas ② 186
5.7.3 Diagramas da regressão ② 188
5.7.4 Salvando os diagnósticos da regressão ② 189
5.7.5 Opções adicionais ② 190
5.8 Interpretando a regressão múltipla ② 191
5.8.1 Descritivas ② 192
5.8.2 Resumo do modelo ② 193
5.8.3 Parâmetros do modelo ② 197
5.8.4 Variáveis excluídas ② 201
5.8.5 Avaliando as hipóteses de não-multicolinearidade ② 202
5.8.6 Diagnósticos por casos (Casewise) ② 204
5.8.7 Conferindo as hipóteses ② 208
5.9 Como relatar a regressão múltipla ② 212
5.10 Previsores categóricos e a regressão múltipla ③ 212
5.10.1 Variáveis dummy ③ 213
5.10.2 Saída do SPSS para variáveis auxiliares (dummy) ③ 215
5.11 O que descobrimos sobre estatística? ① 219
5.12 Termos-chave que descobrimos 219
5.13 Tarefas do Alex Esperto 220
5.14 Leituras complementares 220

6 REGRESSÃO LOGÍSTICA 221
6.1 O que você vai aprender neste capítulo? ① 221
6.2 Pressupostos da regressão logística ① 221
6.3 Quais são os princípios por trás da regressão logística? ③ 222
6.3.1 Avaliando o modelo: a estatística conhecidos de verossimilhança-log ③ 223
6.3.2 Avaliando o modelo: R *e* R^2 ③ 224
6.3.3 Avaliando a contribuição dos previsores: a estatística de Wald ② 225
6.3.4 Exp b ③ 226
6.3.5 Métodos de regressão logística ② 227
6.3.5.1 O método da entrada forçada 227
6.3.5.2 Métodos passo a passo (*stepwise*) 227

Material com direitos autorais

6.3.5.3 Como selecionamos um método? ② 228
6.4 Executando a análise: um exemplo de pesquisa ② 228
6.4.1 A análise principal ② 229
6.4.2 Métodos de regressão ② 230
6.4.3 Previsores categóricos ② 230
6.4.4 Obtendo resíduos ② 231
6.4.5 Opções adicionais ② 232
6.5 Interpretando a regressão logística ② 233
6.5.1 O modelo inicial ② 233
6.5.2 Passo 1: Entendendo falsas crenças ③ 235
6.5.3 Listando probabilidades previstas ② 242
6.5.4 Interpretando os resíduos ② 243
6.5.5 Calculando o tamanho de efeito ② 247
6.6 Como relatar a regressão logística ② 247
6.7 Outro exemplo ② 247
6.7.1 Executando a análise: regressão por entrada em blocos ② 248
6.7.2 Interpretando saídas ③ 250
6.8 Testes para a multicolinearidade ③ 256
6.9 O que pode dar errado ④ 260
6.9.1 Informação incompleta dos previsores ④ 260
6.9.2 Separação completa ④ 261
6.10 O que descobrimos sobre Estatística? ① 262
6.11 Termos-chave que descobrimos 262
6.12 Tarefas do Alex Esperto 262
6.13 Leituras complementares 264

7 COMPARANDO DUAS MÉDIAS 265
7.1 O que você vai aprender neste capítulo? ① 265
7.2 Revisão da pesquisa experimental ① 266
7.2.1 Os dois métodos de coletar dados ① 266
7.2.2 Dois tipos de variação ① 266
7.2.3 Aleatorização ① 268
7.3 Atribuindo dados e apresentando médias com diagramas de barras de erros 269
7.3.1 Diagramas de barras de erros para delineamentos entre grupos ① 271
7.3.2 Diagramas de barras de erros para delineamentos de medidas repetidas ② 274
7.3.2.1 Passo 1: Calcular a média para cada participante ② 275
7.3.2.2 Passo 2: Calcular a média geral ② 275
7.3.2.3 Passo 3: Calcular o fator de ajustamento ② 276
7.3.2.4 Passo 4: Criar os valores ajustados para cada variável ② 276
7.3.2.5 Traçar o diagrama de barras de erros ② 278
7.4 Testando diferenças entre médias: o teste *t* ① 278
7.4.1 Fundamentação do teste t ① 279
7.4.2 Suposições do teste t ① 280
7.5 O teste *t* dependente ① 280
7.5.1 Distribuições amostrais e o erro padrão ① 281
7.5.2 A equação do teste t *dependente explicada* ① 281
7.5.3 O teste t *dependente utilizando o SPSS* ① 283
7.5.4 Saídas do teste t *dependente* ① 284
7.5.5 Calculando o tamanho de efeito ② 286
7.5.6 Relatando o teste t *dependente* ① 286
7.6 O teste *t* independente ① 287
7.6.1 A equação do teste t *independente explicada* ① 287
7.6.2 O teste t *independente utilizando o SPSS* ① 289
7.6.3 Saídas do teste t *independente* ① 290
7.6.4 Calculando o tamanho de efeito ① 292
7.6.5 Relatando o teste t *independente* ① 292
7.7 Entre grupos ou medidas repetidas? ① 293
7.8 O teste *t* como um modelo linear generalizado ② 294
7.9 O que fazer se os dados não forem normalmente distribuídos? ② 295
7.10 O que descobrimos sobre Estatística? ① 295
7.11 Termos-chave que descobrimos 296
7.12 Tarefas do Alex Esperto 296
7.13 Leituras complementares 297

Material com direitos autorais

8 COMPARANDO VÁRIAS MÉDIAS: ANOVA (MLG 1) 298
8.1 O que você vai aprender neste capítulo? ① 298
8.2 A teoria por trás da ANOVA ② 298
8.2.1 Taxas de erros infladas ② 298
8.2.2 ANOVA como regressão ② 299
8.2.3 A lógica da razão F ② 303
8.2.4 Soma dos quadrados Total (SS_T) ② 305
8.2.5 Soma dos Quadrados do Modelo (SS_M) ② 306
8.2.6 Soma dos Quadrados dos Resíduos (SS_R) ② 307
8.2.7 Médias ao quadrado ② 308
8.2.8 A razão F ② 308
8.2.9 Suposições da ANOVA ② 309
8.2.10 Contrastes planejados ② 310
8.2.10.1 Escolhendo que contrastes realizar ② 310
8.2.10.2 Definindo contrastes ponderados ③ 314
8.2.10.3 Comparações não-ortogonais ② 319
8.2.10.4 Contrastes padrões ② 320
8.2.10.5 Contrastes polinomiais: análise da tendência ② 321
8.2.11 Procedimentos *post hoc* ② 322
8.2.11.1 Procedimentos *post hoc* e erros do Tipo I (α) e Tipo II (β) ② 323
8.2.11.2 Procedimentos *post hoc* e violações das hipóteses do teste ② 323
8.2.11.3 Resumo dos Procedimentos *post hoc* ② 324
8.3 Realizando a ANOVA de um fator (One-Way) no SPSS ② 324
8.3.1 Comparações planejadas utilizando o SPSS ② 325
8.3.2 Testes post hoc no SPSS ② 327
8.3.3 Opções ② 328
8.4 Saídas da ANOVA de um fator (*One Way*) ② 328
8.4.1 Saídas para a análise principal ② 328
8.4.2 Saídas para as comparações planejadas ② 333
8.4.3 Saídas para os testes post hoc ② 334
8.5 Calculando o tamanho de efeito ② 338
8.6 Relatando resultados da ANOVA independente de um fator ② 339
8.7 Violações das hipóteses da ANOVA independente de um fator ② 339
8.8 O que descobrimos sobre Estatística? ① 340
8.9 Termos-chave que descobrimos 340
8.10 Tarefas do Alex Esperto 340
8.11 Leituras complementares 341

9 ANÁLISE DE COVARIÂNCIA, ANCOVA (MLG 2) 343
9.1 O que você vai aprender neste capítulo? ② 343
9.2 O que é ANCOVA ② 343
9.3 Executando a ANCOVA no SPSS ② 344
9.3.1 Atribuindo dados ① 344
9.3.2 Análise principal ② 345
9.3.3 Contrastes e outras opções ② 345
9.4 Interpretando as saídas da ANCOVA ② 347
9.4.1 Análise principal ② 347
9.4.2 Contrastes ② 351
9.4.3 Interpretando as covariáveis ② 352
9.5 ANCOVA realizada como uma Regressão Múltipla ② 353
9.6 Suposições adicionais da ANCOVA ③ 357
9.6.1 Homogeneidade dos parâmetros da regressão ③ 357
9.6.2 Testando a homogeneidade dos parâmetros da regressão no SPSS ③ 359
9.7 Calculando o tamanho de efeito ② 361
9.8 Relatando resultados ② 361
9.9 O que descobrimos sobre estatística? ② 362
9.10 Termos-chave que descobrimos 362
9.11 Tarefas do Alex esperto 362
9.12 Leituras complementares 363

10 ANOVA FATORIAL (MLG 3) 364
10.1 O que você vai aprender neste capítulo? ② 364
10.2 Teoria da ANOVA fatorial (entre grupos) ② 364
10.2.1 Delineamentos fatoriais ② 364
10.2.2 Um exemplo com duas variáveis independentes ② 365
10.2.3 A Soma dos Quadrados (SS_T) ② 366

Material com direitos autorais

10.2.4 A Soma dos Quadrados do Modelo (SS_M) ② 367
10.2.4.1 O efeito principal do Gênero (SS_A) ② 367
10.2.4.2 O efeito principal do Álcool (SS_B) ② 368
10.2.4.3 O efeito interação ($SS_{A\times B}$) ② 368
10.2.5 A Soma dos Quadrados dos Resíduos (SS_R) ② 368
10.2.6 As razões F ② 369
10.3 ANOVA fatorial utilizando o SPSS ② 370
10.3.1 Entrando com dados e utilizando 370
10.3.2 Modelos customizados ② 371
10.3.3 Ilustrando interações ② 372
10.3.4 Contrastes ② 373
10.3.5 Testes post hoc ② 373
10.3.6 Opções ② 374
10.4 Saídas da ANOVA fatorial ② 375
10.4.1 Saídas para as análises preliminares ② 375
10.4.2 O teste de Levene ② 375
10.4.3 A tabela principal da ANOVA ② 376
10.4.4 Contrastes ② 378
10.4.5 Análise post hoc ② 380
10.5 Interpretando gráficos de interações ② 383
10.6 Calculando o tamanho de efeito ③ 387
10.7 Relatando os resultados da ANOVA de dois fatores (two way) ② 388
10.8 ANOVA de dois fatores como Regressão ③ 389
10.9 O que descobrimos sobre Estatística? ② 392
10.10 Termos-chave que descobrimos 393
10.11 Tarefas do Alex Esperto 393
10.12 Leituras complementares 394

11 DELINEAMENTOS DE MEDIDAS REPETIDAS (MLG 4) 395
11.1 O que você vai aprender neste capítulo? ② 395
11.2 Introdução aos Delineamentos de Medidas Repetidas ② 395
11.2.1 A suposição de esfericidade ② 396
11.2.2 Como a esfericidade é medida? ② 396
11.2.3 Interpretando as violações da hipótese de esfericidade ② 397
11.2.4 Qual é o efeito de violar a hipótese de esfericidade ③ 397
11.2.5 O que fazer se a hipótese de esfericidade for violada? ② 398
11.3 Teoria da ANOVA de Medidas Repetidas de Um Fator ② 399
11.3.1 A Soma dos Quadrados Total (SS_T) ② 401
11.3.2 A variação dentre participantes (SS_W) ② 401
11.3.3 A Soma dos Quadrados do Modelo (SS_M) ② 402
11.3.4 A Soma dos Quadrados dos Resíduos (SS_R) ② 403
11.3.5 A média dos quadrados ② 403
11.3.6 A razão F ② 403
11.4 ANOVA de Medidas Repetidas de um Fator utilizando o SPSS ② 404
11.4.1 A análise principal ② 404
11.4.2 Definindo contrastes para medidas repetidas ② 405
11.4.3 Testes post hoc e opções adicionais ③ 406
11.5 Saídas para a ANOVA de Medidas Repetidas de Um Fator ② 408
11.5.1 Descritivas e outros diagnósticos ① 408
11.5.2 Interpretando e corrigindo a esfericidade ② 409
11.5.3 A ANOVA principal ② 410
11.5.4 Contrastes ② 413
11.5.5 Testes post hoc ② 414
11.6 Tamanho de efeito para ANOVA de Medidas Repetidas ③ 415
11.7 Relatando a ANOVA de Medidas Repetidas com Um Fator ② 417
11.8 Medidas Repetidas com várias Variáveis Independentes ② 417
11.8.1 A análise principal ② 418
11.8.2 Contrastes ② 421
11.8.3 Ilustrando interações ② 422
11.8.4 Outras opções ② 422
11.9 Saídas para a ANOVA Fatorial de Medidas Repetidas ② 426
11.9.1 Descritivas e análise principal ② 426
11.9.2 O efeito da bebida ② 427
11.9.3 O efeito da imagem ② 430
11.9.4 O efeito interação (bebida × imagem) ② 431

Material com direitos autorais

11.9.5 Contrastes para variáveis de medidas repetidas ② 433
11.9.5.1 Cerveja *versus* água, percepção positiva *versus* neutra ② 434
11.9.5.2 Cerveja *versus* Água, Percepção Negativa *versus* Neutra ② 435
11.9.5.3 Vinho *versus* água, percepção positiva *versus* neutra ② 435
11.9.5.4 Vinho *versus* água, percepção negativa *versus* neutra ② 436
11.9.5.5 Limitações desses contrastes ② 437
11.10 Tamanho de efeito para a ANOVA Fatorial de Medidas Repetidas ③ 437
11.11 Relatando os resultados da ANOVA Fatorial de Medidas Repetidas ② 438
11.12 O que descobrimos sobre Estatística? ② 439
11.13 Termos-chave que descobrimos 439
11.14 Tarefas do Alex Esperto 440
11.15 Leituras complementares 440

12 ANOVA COM DELINEAMENTOS MISTOS (MLG 5) 441
12.1 O que você vai aprender neste capítulo? ② 441
12.2 O que homens e mulheres procuram em um companheiro? ② 441
12.3 ANOVA mista com o SPSS ② 442
12.3.1 A análise principal ② 442
12.3.2 Outras opções ② 445
12.4 Saídas para a ANOVA mista fatorial: Análise principal ③ 446
12.4.1 O efeito de gênero ② 450
12.4.2 O efeito da aparência ② 451
12.4.3 O efeito do carisma ② 454
12.4.4 A interação entre gênero e aparência ② 454
12.4.4.1 Interação 1 aparência × gênero: atrativo × mediano, homem × mulher ② 456
12.4.4.2 Interação 2 aparência × gênero: feio × mediano, homem × mulher ② 456
12.4.5 A interação entre gênero e carisma ② 456
12.4.5.1 Interação 1 carisma × gênero: alta × algum carisma, homem × mulher ② 457
12.4.5.2 Interação 2 carisma × gênero: sem carisma *versus* algum carisma, homem *versus* mulher ② 457
12.4.6 A interação entre atraente e carismático ② 458
12.4.6.1 Interação 1 aparência × carisma: atraente *versus* mediano, carisma alto *versus* algum carisma ② 459
12.4.6.2 Interação 2 aparência × carisma: atraente *versus* mediano, sem carisma *versus* algum carisma ② 460
12.4.6.3 Interação 3 aparência × carisma: feio *versus* mediano, carisma alto *versus* algum carisma ② 461
12.4.6.4 Interação 4 aparência × carisma: feio *versus* mediano, sem carisma *versus* algum carisma ② 461
12.4.7 A interação entre aparência, carisma e gênero ③ 462
12.4.7.1 Interação 1 aparência × carisma × gênero: atraente *versus* mediano, carisma alto *versus* algum carisma, homem *versus* mulher ③ 464
12.4.7.2 Interação 2 aparência × carisma × gênero: atraente *versus* mediano, sem carisma *versus* algum carisma, homem *versus* mulher ③ 464
12.4.7.3 Interação 3 aparência × carisma × gênero: feio *versus* mediano, carisma alto *versus* algum carisma, homem *versus* mulher ③ 465
12.4.7.4 Interação 4 aparência × carisma × gênero: atraente *versus* mediano, sem carisma *versus* algum carisma, homem *versus* mulher ③ 466
12.4.8 Conclusões ③ 467
12.5 Calculando o tamanho de efeito ③ 468
12.6 Relatando os resultados da ANOVA Mista ② 469

Material com direitos autorais

12.7 O que descobrimos sobre Estatística? ② 471
12.8 Termos-chave que descobrimos 472
12.9 Tarefas do Alex Esperto 472
12.10 Leituras complementares 473

13 TESTES NÃO-PARAMÉTRICOS 474
13.1 O que você vai aprender neste capítulo? ① 474
13.2 Comparando duas condições independentes: o teste da soma dos postos de Wilcoxon e o Teste de Mann-Whitney ① 475
13.2.1 Teoria ② 475
13.2.2 Atribuindo dados e análise condicional ① 478
13.2.3 Executando a análise ① 478
13.2.4 Saídas para o teste de Mann-Whitney ① 481
13.2.5 Calculando o tamanho de efeito ② 482
13.2.6 Escrevendo os resultados ① 483
13.3 Comparando duas condições relacionadas: o teste dos Sinais de Wilcoxon ① 484
13.3.1 Teoria do teste dos postos com sinais de Wilcoxon ② 484
13.3.2 Executando a análise ① 486
13.3.3 Saídas para o grupo do ecstasy ① 488
13.3.4 Saídas para o grupo do álcool ① 489
13.3.5 Calculando o tamanho de efeito ② 490
13.3.6 Escrevendo e interpretando os resultados ① 490
13.4 Diferenças entre vários grupos independentes: o teste de Kruskal-Wallis ① 490
13.4.1 Teoria do teste de Kruskal-Wallis ② 492
13.4.2 Atribuindo dados e fazendo análise condicional ① 493
13.4.3 Executando o teste de Kruskal-Wallis no SPSS ① 493
13.4.4 Saídas para o teste de Kruskal-Wallis ① 495
13.4.5 Análise post hoc para o teste de Kruskal-Wallis ② 496
13.4.6 Testes para a tendência: o teste de Jonckheere-Terpstra ② 500
13.4.7 Calculando o tamanho de efeito ② 502
13.4.8 Escrevendo e interpretando os resultados ① 502
13.5 Diferenças entre vários grupos relacionados: a ANOVA de Friedman ① 503
13.5.1 Teoria da ANOVA de Friedman ② 504
13.5.2 Atribuindo dados e fazendo análise condicional ① 505
13.5.3 Executando a ANOVA de Friedman no SPSS ① 505
13.5.4 Saídas para a ANOVA de Friedman ① 506
13.5.5 Análise post hoc para a ANOVA de Friedman ② 508
13.5.6 Calculando o tamanho de efeito ② 510
13.5.7 Escrevendo e interpretando os resultados ① 510
13.6 O que descobrimos sobre Estatística? ① 511
13.7 Termos-chave que descobrimos 512
13.8 Tarefas do Alex Esperto 512
13.9 Leituras complementares 513

14 ANÁLISE DE VARIÂNCIA MULTIVARIADA (MANOVA) 514
14.1 O que você vai aprender neste capítulo? ② 514
14.2 Introdução: Semelhanças e Diferenças com a ANOVA ② 514
14.2.1 Advertência ② 515
14.2.2 Controvérsias ② 515
14.2.3 O exemplo do capítulo ② 516
14.3 Teoria da MANOVA ③ 517
14.3.1 Introdução às matrizes ③ 518
14.3.2 Algumas matrizes importantes e suas funções ③ 518
14.3.3 Calculando a MANOVA manualmente: um exemplo ③ 519
14.3.3.1 ANOVA univariada para VD 1 (Ações) ② 519
14.3.3.2 ANOVA univariada para VD 2 (pensamentos) ② 520
14.3.3.3 O relacionamento entre VDs: produtos cruzados ② 521
14.3.3.4 A matriz SSCP TOTAL (T) ③ 523
14.3.3.5 A matriz SSCP Residual (E) ③ 524
14.3.3.6 A matriz SSCP do modelo (H) ③ 525

Material com direitos autorais

14.3.4 Princípios do teste estatístico MANOVA ④ 525
14.3.4.1 A função discriminante ④ 526
14.3.4.2 Traço de Pillai-Bartlett (V) ④ 528
14.3.4.3 T^2 de Hotelling ④ 529
14.3.4.4 Lambda de Wilks (Λ) ④ 529
14.3.4.5 Maior raiz de Roy ④ 529
14.4 Hipóteses da MANOVA ③ 530
14.4.1 Verificando as hipóteses ③ 530
14.4.2 Escolhendo o teste estatístico ③ 531
14.4.3 Análise adicional ③ 531
14.5 MANOVA no SPSS ② 532
14.5.1 A análise principal ② 532
14.5.2 Comparações múltiplas na MANOVA ② 533
14.5.3 Opções adicionais ③ 533
14.6 Saídas da MANOVA ③ 534
14.6.1 Análise preliminar e checagem de hipóteses ③ 534
14.6.2 Teste estatístico MANOVA ③ 536
14.6.3 Teste estatístico Univariado ② 536
14.6.4 Matrizes SSCP ③ 538
14.6.5 Contrastes ③ 540
14.7 MANOVA adicional com Análise Discriminante ③ 542
14.8 Saídas da Análise Discriminante ④ 545
14.9 Algumas considerações finais ④ 548
14.9.1 A interpretação final ④ 548
14.9.2 ANOVA univariada ou Análise discriminante? ③ 549
14.10 O que descobrimos sobre Estatística? ② 549
14.11 Termos-chave que descobrimos 551
14.12 Tarefas do Alex Esperto 551
14.13 Leituras complementares 552

15 ANÁLISE DE FATORES EXPLORATÓRIA 553
15.1 O que você vai aprender neste capítulo? ② 553
15.2 Fatores ② 553
15.2.1 Representação gráfica dos fatores ② 555
15.2.2 Representação matemática dos fatores ② 556
15.2.3 Escores dos fatores ② 558
15.2.3.1 O método da regressão ④ 559
15.2.3.2 Outros métodos ② 560
15.2.3.3 Usos dos escores dos fatores ② 560
15.3 Descobrindo fatores ② 561
15.3.1 Escolhendo um método ② 561
15.3.2 Comunalidades ② 562
15.3.3 Análise de fatores versus *Análise de Componentes Principais ② 562*
15.3.4 A teoria por trás da Análise de Componentes Principais ③ 563
15.3.5 Extração de fatores: autovalores e o diagrama de inclinação ② 564
15.3.6 Melhorando a interpretação: a rotação de fatores ③ 566
15.3.6.1 Escolhendo um método de rotação dos fatores ③ 568
15.3.6.2 Importância da carga dos fatores ② 568
15.4 Exemplo de pesquisa ② 569
15.4.1 Primeiras considerações ② 570
15.4.1.1 Tamanho da amostra ② 570
15.4.1.2 Vasculhando os dados ② 571
15.5 Executando a análise ② 572
15.5.1 Extração de fatores no SPSS ② 573
15.5.2 Rotação ② 575
15.5.3 Escores ② 576
15.5.4 Opções ② 576
15.6 Interpretando saídas do SPSS ② 577
15.6.1 Análise preliminar ② 577
15.6.2 Extração de fatores ② 581
15.6.3 Rotação de fatores ② 587
15.6.3.1 Rotação ortogonal (Varimax) ② 587
15.6.3.2 Rotação oblíqua ② 589
15.6.4 Escores do fator ② 592
15.6.5 Resumo ② 593
15.7 Análise de confiabilidade ② 593
15.7.1 Medidas de confiabilidade ③ 593
15.7.2 Interpretando o α de Cronbach (advertências) ② 594
15.7.3 Análise de confiabilidade no SPSS ② 597
15.7.4 Interpretando a saída ② 598
15.8 O que descobrimos sobre Estatística? ② 602
15.9 Termos-chave que descobrimos 603
15.10 Tarefas do Alex Esperto 604
15.11 Leituras complementares 604

Material com direitos autorais

16 DADOS CATEGÓRICOS 606
16.1 O que você vai aprender neste capítulo? ① 606
16.2 Teoria da Análise de Dados Categóricos ① 606
16.2.1 Teste qui-quadrado de Pearson ① 607
16.2.2 A razão de verossimilhança ② 608
16.2.3 Correção de Yates ② 609
16.3 Hipóteses do teste qui-quadrado ① 609
16.4 Realizando o teste qui-quadrado no SPSS ① 610
16.4.1 Entrando com os dados: escores brutos ① 610
16.4.2 Entrando com os dados: ponderando casos ① 610
16.4.3 Executando a análise ① 611
16.4.4 Saída para o teste qui-quadrado ① 613
16.4.5 Calculando o tamanho de efeito ② 616
16.4.6 Relatando os resultados do qui-quadrado ① 617
16.5 Diversas variáveis categóricas: Análise Loglinear ③ 618
16.5.1 Qui-quadrado como regressão ④ 618
16.5.2 Análise Loglinear ③ 623
16.6 Hipóteses na Análise Loglinear ② 625
16.7 Análise Loglinear utilizando o SPSS ② 625
16.7.1 Considerações iniciais ② 625
16.7.2 A Análise Loglinear ② 626
16.8 Saídas da Análise Loglinear ③ 629
16.9 Continuação da Análise Loglinear ② 635
16.10 Tamanhos de efeito na Análise Loglinear ③ 636
16.11 Relatando os resultados da Análise Loglinear ③ 637
16.12 O que descobrimos sobre Estatística? ① 637
16.13 Termos-chave que descobrimos 638
16.14 Tarefas do Alex Esperto 638
16.15 Leituras complementares 639

17 EPÍLOGO 640

GLOSSÁRIO 641

APÊNDICE 663
A.1 Tabela da distribuição normal padrão 663
A.2 Valores críticos da distribuição t 667
A.3 Valores críticos da distribuição F 668
A.4 Valores críticos da distribuição qui-quadrado 671
A.5 O teste F de Welch 671
A.6 Calculando efeitos simples 671
A.7 Teste da tendência de Jonckheere 671
A.8 Capítulo 14 671
A.9 Cálculos dos coeficientes dos escores dos fatores 671
Referências 673
Índice 679

Material com direitos autorais

1

TUDO O QUE VOCÊ SEMPRE QUIS SABER SOBRE ESTATÍSTICA (BEM, QUASE TUDO)

1.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ①

Sei que muitas pessoas usarão este livro para consultar somente os capítulos que descrevem os testes estatísticos que elas descobriram de uma hora para outra que precisam utilizar. No entanto, para que esses capítulos façam sentido, talvez seja útil saber algumas coisas sobre estatística. Este capítulo é uma tentativa de fornecer um breve panorama de alguns conceitos estatísticos importantes, como o de utilizar modelos estatísticos para responder questões científicas.

1.2 CONSTRUINDO MODELOS ESTATÍSTICOS ①

Nas ciências sociais, normalmente estamos interessados em descobrir algo sobre um fenômeno que acreditamos que realmente exista (um fenômeno do mundo real). Esses fenômenos reais podem ser qualquer coisa desde o comportamento da taxa de juros no mercado econômico até comportamento de universitários numa festa de final de semestre. Qualquer que seja o fenômeno que desejamos explicar, procuramos explicá-lo coletando dados do mundo real e então utilizando esses dados para tirar conclusões sobre o que está sendo estudado. Como estatísticos, nosso trabalho é pegar os dados disponíveis e utilizá-los de uma forma apropriada e isso, frequentemente, envolve construir modelos estatísticos do fenômeno de interesse.

A razão para construirmos modelos estatísticos dos dados do mundo real é melhor explicada por analogia. Imagine que um engenheiro quer construir uma ponte sobre um rio. Esse engenheiro seria bem bobo se construísse qualquer ponte com materiais ultrapassados porque ela teria grandes chances de cair. Em vez disso, ele coleta dados do mundo real: olha para as pontes do mundo real e observa de que material elas são feitas, que estruturas foram utilizadas (ele pode inclusive coletar dados sobre os defeitos que elas apresentam!). Ele, então, usa essa informação para construir um modelo. O engenheiro constrói um modelo, em escala, da ponte do mundo real porque é impraticável, além de caro, construir a ponte real. O modelo será diferente da realidade em muitos aspectos – para começar, ele será menor – mas o engenheiro irá tentar construir o modelo que melhor se adapte ao seu interesse com base nos dados disponíveis. Uma vez que o modelo foi construído, ele pode ser usado para prever coisas sobre o mundo real: por

Material com direitos autorais

exemplo, o engenheiro poderá testar se a ponte irá suportar fortes ventos colocando o modelo num túnel de vento. Obviamente, é necessário que o modelo seja uma representação precisa do mundo real. Cientistas sociais fazem a mesma coisa que os engenheiros: constroem modelos de processos do mundo real na tentativa de prever como esses processos operam sob certas circunstâncias. Não temos acesso direto a esses processos, assim, coletamos dados que representam o processo e usamos esses dados para construir modelos estatísticos (reduzimos o processo a um modelo estatístico). Depois, usamos esse modelo estatístico para fazer previsões sobre o fenômeno do mundo real. Assim como o engenheiro, queremos que nosso modelo seja o mais preciso possível, para que possamos garantir que as previsões que fizermos serão também precisas. Entretanto, diferente dos engenheiros, não temos acesso a situações do mundo real e, desse modo, podemos somente *inferir* coisas sobre processos psicológicos, sociais ou econômicos baseados nos modelos que construímos. Se quisermos que nossas deduções sejam precisas, o modelo estatístico que construímos deve representar os dados que coletamos (os *dados observados*) sempre que possível. O grau com que o modelo estatístico representa os dados coletados é conhecido como *aderência* do modelo e esse é um termo que você irá ver frequentemente.

A Figura 1.1 ilustra os tipos de modelos que um engenheiro pode construir para representar a ponte do mundo real que ele quer criar. O primeiro modelo (a) é uma excelente representação de uma situação do mundo real e é conhecida como uma *boa aderência* (isto é, existem algumas diferenças, mas o modelo é uma boa réplica da realidade). Se esse modelo for utilizado para fazer previsões sobre o mundo real, o engenheiro pode confiar que essas previsões serão precisas, porque o modelo representa bem a realidade. Assim, se o modelo desabar com um vento forte, haverá uma boa chance de que a ponte real desabará também. O segundo modelo (b) tem algumas semelhanças com o mundo real: o modelo inclui algumas características básicas de estrutura, mas

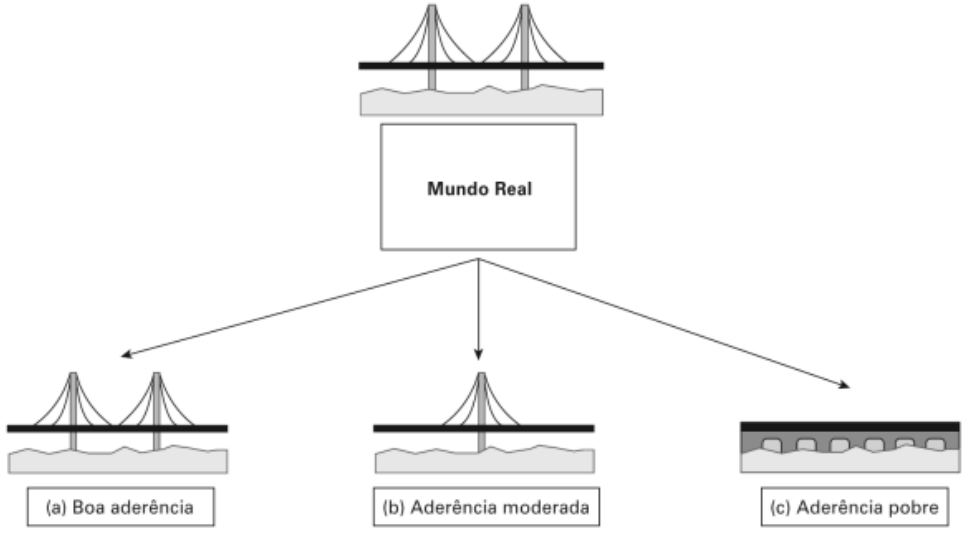


Figura 1.1 Ajustando modelos ao mundo real (veja o texto para detalhes).

Material com direitos autorais

há grandes diferenças da ponte do mundo real (isto é, a ausência de uma das torres de suporte). Chamamos isso de *aderência moderada* (isto é, há algumas diferenças entre o modelo e os dados, mas há também grandes similaridades). Se o engenheiro usar esse modelo para fazer previsões sobre o mundo real, essas previsões poderão ser imprecisas e possivelmente catastróficas (por exemplo, se a ponte desabar com ventos fortes isso será devido a ausência da segunda torre de suporte). Portanto, o uso desse modelo resulta em previsões nas quais podemos confiar parcialmente, mas não totalmente. O modelo final (c) é completamente diferente da situação do mundo real. Esse modelo não possui semelhanças estruturais com a ponte real e pode ser chamado de modelo de pouca aderência (na verdade, pode ser mais precisamente descrito como uma aderência abismal!). Assim, qualquer previsão baseada nesse modelo provavelmente será imprecisa. Estendendo essa analogia para as ciências sociais, podemos dizer que é importante quando um modelo estatístico adere bem a um conjunto de dados. Se o nosso modelo não aderir bem aos dados observados, nossas previsões serão igualmente pobres.

1.3 POPULAÇÕES E AMOSTRAS ①

Como pesquisadores, estamos interessados em encontrar resultados que se apliquem a toda uma **população** de pessoas ou coisas. Por exemplo, psicólogos querem descobrir processos que ocorrem em todos humanos, biólogos podem estar interessados em processos que ocorrem em todas as células, economistas querem construir modelos que se aplicam a todos os salários, e assim por diante. A população pode ser geral (todos os seres humanos) ou muito pequena (todos os gatos machos de pelo castanho-avermelhado chamados Bob), mas em ambos os casos os cientistas raramente, senão nunca, terão acesso a cada membro de uma população. Psicólogos não podem coletar dados de cada ser humano e ecologistas não podem observar cada gato macho de pelo castanho-avermelhado chamado Bob. Portanto, coletamos dados de um pequeno subconjunto de uma população (chamado de *amostra*) e usamos essas informações para inferir coisas sobre toda a população. O engenheiro construtor de pontes não pode fazer um modelo de tamanho real da ponte que ele quer construir, assim, ele constrói um modelo em escala menor e testa esse modelo sob várias condições. Dos resultados obtidos do modelo em escala menor, o engenheiro infere coisas sobre como a ponte de tamanho real irá responder. O modelo em escala menor poderá responder diferentemente da versão em escala real, mas, quanto maior o modelo, maior a probabilidade de ele se comportar como a ponte de tamanho real. Essa metáfora pode ser estendida a cientistas sociais. Nunca temos acesso à população inteira (a ponte de tamanho real), assim, coletamos pequenas amostras (a ponte em escala menor) e usamos o comportamento dentro da amostra para inferir coisas sobre o comportamento da população. Quanto maior a amostra, maior a probabilidade de ela refletir a população inteira. Se pegarmos muitas amostras aleatórias da população, cada uma dessas amostras fornecerá resultados ligeiramente diferentes. Entretanto, em média, resultados de grandes amostras deverão ser bastante similares.

1.4 MODELOS ESTATÍSTICOS SIMPLES ①

1.4.1 A média, a soma dos quadrados, a variância e o desvio padrão ①

Um dos modelos mais simples usados em estatística é a média. Alguns de vocês podem estranhar pensar na média como um modelo, mas, na verdade, é porque ela representa um resumo dos dados. A **média** é um valor hipotético que pode ser calculado para qualquer conjunto de dados; ela não precisa ser um valor realmente observado no conjunto de dados. Por exemplo, se pegarmos cinco professores de estatística e contarmos o número de amigos que eles têm, poderemos encontrar os seguintes dados: 1, 2, 3, 3 e 4. Se pegarmos

Material com direitos autorais

a média do número de amigos, isso poderá ser calculado adicionando os valores obtidos e dividindo pelo número de valores contados: (1 + 2 + 3 + 3 + 4)/5 = 2,6. Sabemos que é impossível ter 2,6 amigos (a não ser que você corte alguém com uma serra e seja amigo do seu braço), assim, a média é um valor *hipotético*. Portanto, a média é um modelo criado para resumir nossos dados. Agora, podemos determinar se esse modelo é preciso verificando quão diferente os nossos dados reais são do modelo que criamos. Uma maneira de fazer isso é olhar a diferença entre os dados que observamos e o modelo ajustado. A Figura 1.2 mostra o número de amigos que cada professor de estatística tem e, também, o número médio que anteriormente calculamos. A linha representando a média é o nosso modelo e os círculos são os dados observados. O diagrama tem, também, uma série de linhas verticais que conectam cada valor observado ao valor da média. Essas linhas representam os **desvios** entre os dados observados e o nosso modelo e podem ser pensadas como o erro do modelo. Podemos calcular a magnitude desses desvios simplesmente subtraindo a média ($\bar{x}$) de cada um dos valores observados (x_i)[1]. Por exemplo, o professor 1 tem somente 1 amigo, portanto, a diferença é: $x_i - \bar{x} = 1 - 2{,}6 = -1{,}6$. Você deve notar que o desvio é um número negativo e ele representa o fato de que nosso modelo superestimou a popularidade desse professor: ele prevê que o professor terá 2,6 amigos quando, na verdade, ele tem somente 1 amigo (sorte para ele!). Agora, como podemos usar esses desvios para estimar a precisão do modelo? Uma possibilidade é somar os desvios (isso fornecerá uma estimativa do erro total). Se fizermos isso, acharemos:

$$\begin{aligned}\text{Erro total} &= \text{soma dos desvios} \\ &= \sum(x_i - \bar{x}) = (-1{,}6) + (-0{,}6) \\ &\quad + (0{,}4) + (0{,}4) + (1{,}4) = 0\end{aligned}$$

O resultado diz que não há um erro total entre nosso modelo e os dados observados,

[1] O x_i simplesmente se refere ao escore observado para a i-ésima pessoa (assim, pode ser trocado por um número que representa um determinado indivíduo). Para esses dados: o professor 1, $x_i = x_1 = 1$; o professor 3, $x_i = x_3 = 3$; o professor 5, $x_i = x_5 = 4$.

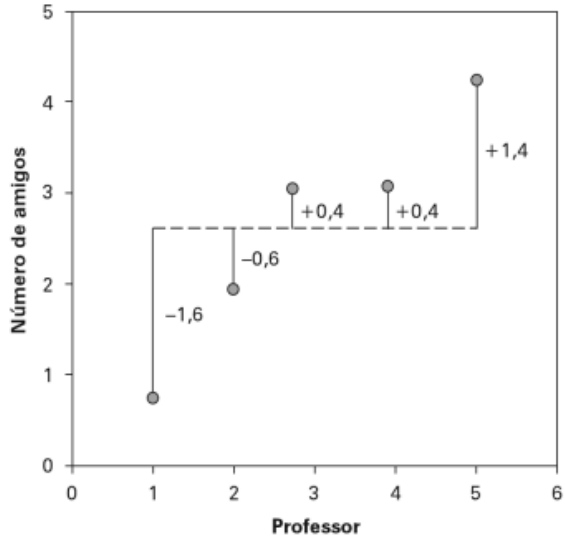


Figura 1.2 Gráfico mostrando as diferenças entre o número observado de amigos que cada professor de estatística tem e o número médio de amigos (linha tracejada).

Material com direitos autorais

assim, a média é uma representação perfeita dos dados. Mas isso obviamente não é verdade: houve erros, alguns deles foram positivos e alguns negativos e eles simplesmente cancelaram uns aos outros. Devemos evitar o problema do erro direcionado (isto é, positivo ou negativo) e uma maneira matemática de fazer isso é elevar cada erro ao quadrado[2], ou seja, multiplicar cada erro por ele mesmo. Assim, em vez de calcularmos a soma dos erros, calculamos a soma dos quadrados dos erros. Neste exemplo:

$$\begin{aligned}\text{Soma dos erros ao quadrado (SS)} &= \sum (x_i - \bar{x})\,(x_i - \bar{x}) \\ &= (-1{,}6)^2 + (-0{,}6)^2 \\ &\quad + (0{,}4)^2 + (0{,}4)^2 + (1{,}4)^2 \\ &= 2{,}56 + 0{,}36 + 0{,}16 \\ &\quad + 0{,}16 + 1{,}96 = 5{,}20\end{aligned}$$

A **soma dos erros ao quadrado** (SS) é uma boa medida da acurácia do nosso modelo. Contudo, é óbvio que a soma dos erros ao quadrado depende do total de dados que foram coletados – quanto mais dados, maior o SS. Para solucionar esse problema, calculamos a média dos erros dividindo o SS pelo número de observações (N). Se estivermos somente interessados na média do erro para a amostra, podemos dividir apenas por N. Entretanto, geralmente estamos interessados em usar o erro na amostra para estimar o erro na população e, assim, dividimos o SS pelo número de observações menos 1 (a explicação está no Quadro 8.2). Essa medida é conhecida como **variância** e você a usará muito.

$$\begin{aligned}\text{Variância} = s^2 &= \frac{\text{SS}}{N-1} \\ &= \frac{\sum (x_i - \bar{x})^2}{N-1} = \frac{5{,}20}{4} = 1{,}30\end{aligned}$$

A variância é, portanto, a média do erro entre a média e as observações feitas (e é a medida de como o modelo corresponde aos dados reais). Existe um problema da variância como medida: ela é expressa em unidades quadradas (porque colocamos cada erro ao quadrado no cálculo). Em nosso exemplo, teremos que dizer que a média do erro em nossos dados (a variância) foi 1,3 amigos ao quadrado. Se já faz pouco sentido falar de 1,3 amigos, faz menos sentido ainda falar de amigos ao quadrado! Por essa razão, geralmente tiramos a raiz quadrada da variância (o que garante que o erro médio será expresso na mesma unidade da variável). Essa medida é conhecida como **desvio padrão** e é simplesmente a raiz quadrada da variância. Neste exemplo o desvio padrão é:

$$\begin{aligned}\text{Desvio padrão} = s &= \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}} \\ &= \sqrt{1{,}30} \\ &= 1{,}14\end{aligned}$$

O desvio padrão é, portanto, uma medida de quão bem a média representa os dados. Pequenos desvios padrões (relativos ao valor da própria média) indicam que pontos de dados estão próximos da média. Um desvio padrão grande (relativo à média) indica que os pontos de dados estão distantes da média (isto é, a média não é uma representação precisa dos dados). Um desvio padrão de 0 significaria que todos os escores são os mesmos. A Figura 1.3 mostra um índice geral (em uma escala de cinco pontos) dos dois professores depois de cada uma das cinco aulas. Ambos os professores tiveram uma média de 2,6 em uma escala de 5 pontos. Entretanto, o primeiro professor tem um desvio padrão de 0,55 (relativamente pequeno se comparado à média). Deve ficar claro no gráfico que os índices para esse professor estavam consistentemente próximos do valor da média. Houve uma pequena flutuação, mas em geral suas aulas não variaram em popularidade. Como tal, a média é uma representação precisa dos seus índices. A média teve uma boa aderência aos dados. O segundo professor, entretanto, teve um desvio padrão de 1,82 (relativamente alto comparado à média). Os índices para esse professor estão claramente mais espalhados em torno da média; isto é, para

[2] Quando você multiplica um número negativo por ele mesmo ele se torna positivo.

Material com direitos autorais

algumas aulas ele recebeu índices muito altos e para outras seus índices foram horríveis. Portanto, a média não é uma representação tão precisa da sua performance porque houve muita variação na popularidade das suas aulas. A média mostrou uma aderência pobre aos dados, nesse caso. Esta ilustração deverá deixar claro por que o desvio padrão é uma medida de quão bem a média representa os dados.

A discussão das médias, somas dos quadrados e da variância pode parecer um desvio do ponto inicial, mas, na verdade, a média é provavelmente um dos modelos estatísticos mais simples que pode ser ajustado aos dados. O que eu quero dizer com isso? Bem, tudo em estatística se resume a uma equação:

$$\text{Saída}_i = (\text{Modelo}_i) + \text{erro}_i \qquad (1.1)$$

Isso apenas significa que os dados que observamos podem ser previstos pelo modelo que escolhemos para ajustar os dados mais um erro. Quando eu digo que a média é um modelo estatístico simples, o que quero dizer é que podemos trocar a palavra "modelo" pela palavra "média", na equação (1.1). Se voltarmos ao nosso exemplo envolvendo o número de amigos que os professores de estatística têm e olharmos para o primeiro professor, por exemplo, observamos que ele tinha um amigo e a média de todos os professores foi 2,6. Assim, a equação se torna:

$$\begin{aligned}\text{Saída}_{\text{professor1}} &= \overline{X} + \varepsilon_{\text{professsor1}} \\ 1 &= 2{,}6 + \varepsilon_{\text{professor1}}\end{aligned}$$

Disso podemos entender que o erro é 1 − 2,6, ou −1,6. Se substituirmos esse valor dentro da equação, teremos 1 = 2,6 − 1,6 ou 1 = 1. Embora isso pareça óbvio, vale ter essa equação em mente ao longo do livro, porque você irá descobrir que a maioria das coisas se resume a uma ideia simples!

Da mesma forma, a variância e o desvio padrão ilustram outro conceito fundamental: como a aderência de um modelo pode ser medida. Se estivermos interessados em avaliar quão bem o modelo se adequou aos dados (nesse caso nosso modelo é a média), então geralmente olhamos para os desvios do modelo, avaliamos a soma dos erros ao quadrado e, de forma geral, podemos escrever isso como:

$$\text{Desvio} = \sum(\text{observado} - \text{modelo})^2 \qquad (1.2)$$

Ou seja, avaliamos modelos comparando os dados que observamos ao modelo utilizado e elevamos as diferenças observadas ao qua-

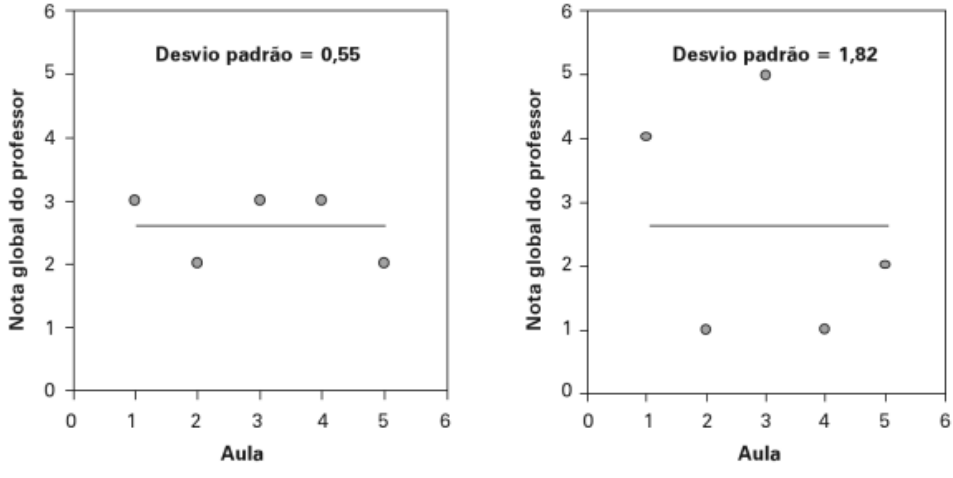


Figura 1.3 Gráfico ilustrando dados que apresentam a mesma medida, mas desvios padrão diferentes.

Material com direitos autorais

drado. Essa ideia fundamental se repetirá ao longo de todo o livro.

1.5 DISTRIBUIÇÕES DE FREQUÊNCIAS ①

Depois de coletar alguns dados, algo muito útil a fazer é um gráfico de quantas vezes cada escore ocorre. Isso é conhecido como **distribuição de frequências ou histograma**, que é simplesmente um gráfico com os valores observados no eixo horizontal, com barras mostrando quantas vezes cada valor ocorreu no conjunto de dados. A distribuição de freqüências pode ser útil para avaliar as propriedades de um conjunto de valores. Olhando para o escore que apresenta a barra mais alta, podemos imediatamente ver a **moda**, que é simplesmente o escore que ocorre mais frequentemente num conjunto de dados.

Distribuições de frequências ocorrem de muitas formas e tamanhos diferentes. É muito importante, portanto, ter algumas descrições gerais para os tipos mais comuns de distribuições. Em um mundo ideal, nossos dados estariam distribuídos simetricamente em volta do centro de todos os escores. Assim, se traçássemos uma linha vertical pelo centro da distribuição, ela deveria ser a mesma em ambos os lados. Isso é conhecido como **distribuição normal** e é caracterizado por uma curva em forma de sino, a qual você já deve conhecer. Essa forma basicamente sugere que a maioria dos escores está em torno do centro da distribuição (assim, as barras maiores no histograma estão em volta do valor central). Também, à medida que nos distanciamos do centro, as barras ficam menores, sugerindo que à medida que os escores começam a se desviar do centro, sua frequência diminui. Afastando-nos ainda mais do centro a freqüência dos nossos escores se torna muito baixa (as barras são muito pequenas). Um exemplo de distribuição normal é mostrado na Figura 1.4.

1.5.1 Propriedades das distribuições de frequências ①

Uma distribuição pode se desviar de uma normal de duas maneiras principais: (1) falta de simetria (chamado de **assimetria**) e (2) achatamento (chamado de **curtose**). Distribuições assimétricas não são parelhas e, em vez disso, os escores mais frequentes (a parte mais alta do gráfico) estão concentrados em um dos lados da escala. Uma distribuição assimétrica pode ser positivamente assimétrica (a maioria dos escores está concentrada à esquerda da escala) ou negativamente assimétrica (a maioria dos pontos está concentrada à direita da escala). A Figura 1.5 mostra exemplos dessas distribuições.

As distribuições também variam no seu achatamento ou curtose. A curtose, apesar de soar como um tipo de doença exótica refere-se ao grau que os escores estão concentrados na cauda da distribuição. Essa característica é avaliada por um grau de achatamento (achatada ou pontiaguda) da distribuição. Uma distribuição platicúrtica é aquela que têm muitos escores nas caudas (chamada de distribuição com cauda pesada) sendo, portanto, bem achatada. Em contraste, as distribuições leptocúrticas são relativamente finas e parecem bem pontudas. Você pode lembrar disso assim: "distribuições leptocúrticas se elevam no ar enquanto que as distribuições platicúrticas ficam achatadas como um platô (Figura 1.6), ou para os surrealistas entre vocês podem lembrar-se delas como "as distribuições platicúrticas são 'boas de bico' e as distribuições leptocúrticas 'compulsivamente roubam coisas'." Idealisticamente queremos que nossos dados sejam distribuídos de forma normal (isto é, nem muito assimétricos, nem muito pontiagudos ou achatados!).

Em uma distribuição normal os valores da assimetria e da curtose são 0* (isto é, a dis-

* N de R. T. De fato, o coeficiente de curtose fornece um valor igual a 3 para a curva normal. Centrado o valor (isto é, subtraindo 3), temos o coeficiente de curtose centrado que é igual a zero para a curva normal. É esse valor ao qual o autor está se referindo.

Material com direitos autorais

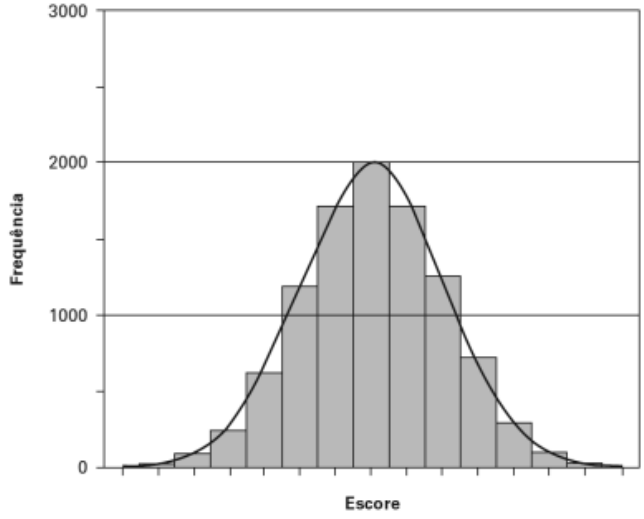


Figura 1.4 Uma distribuição "normal" (a curva mostra a forma idealizada).

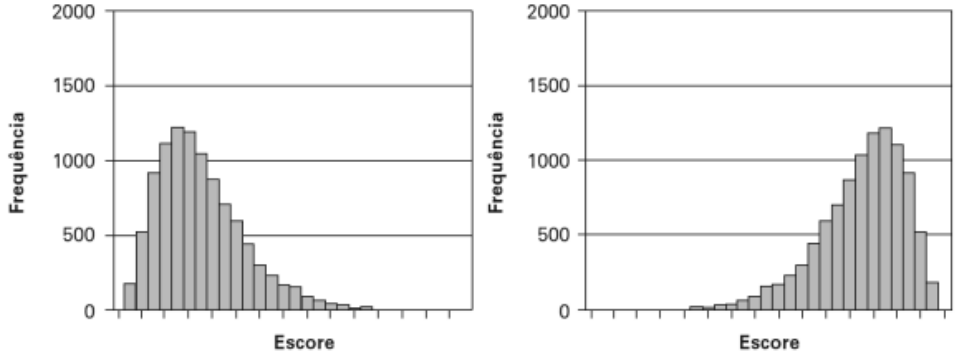


Figura 1.5 Uma distribuição positivamente (esquerda) e negativamente (direita) assimétrica.

tribuição não é muito pontiaguda nem muito achatada e é perfeitamente simétrica). Se a distribuição tem valores de assimetria ou curtose acima ou abaixo de 0, isso indica um desvio da normal (veja o Capítulo 3).

1.5.2 O desvio padrão e a forma da distribuição ①

Assim como fornecem uma ideia da precisão da média como um modelo do nosso conjunto de dados, a variância e o desvio padrão também nos informam sobre a forma da distribuição dos escores. Se a média representa bem os dados, a maioria dos escores irá se concentrar perto da média e o desvio padrão resultante é pequeno quando comparado à média. Quando a média é uma representação ruim dos dados, os escores se espalham mais ao redor da média (lembre-se da Figura 1.3) e o desvio padrão é maior. A Figura 1.7 mostra duas distribuições que tem a mesma média (50), mas desvios padrão diferentes. Uma tem um desvio padrão maior em relação à média

Material com direitos autorais

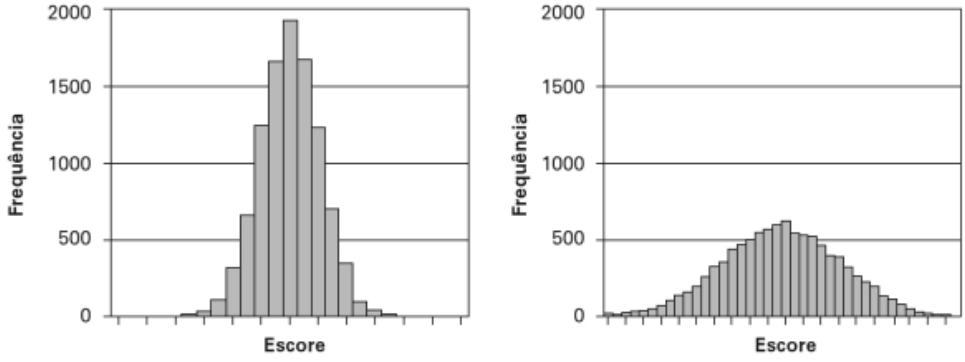


Figura 1.6 Uma distribuição leptocúrtica (esquerda) e platicúrtica (direita).

(DP = 25) e isso resulta em uma distribuição achatada que está mais espalhada, enquanto a outra tem um desvio padrão menor em relação à média (DP = 15), resultando em uma distribuição mais pontiaguda em que os escores próximos da média são mais frequentes, mas escores distantes da média tornam-se muito infrequentes. A mensagem principal é que à medida que o desvio padrão fica maior, a distribuição tende a ficar mais achatada.

1.5.3 O que é a distribuição normal padrão? ①

Outra maneira de olhar as distribuições de frequência é em termos de probabilidade: eles fornecem uma ideia da probabilidade de um dado escore acontecer. Beachy Head é um rochedo grande e ventoso na costa de Sussex (não muito longe de onde eu moro) e tem uma reputação de atrair suicidas, que parecem gostar de se jogar dele (e após meses e meses reescrevendo este livro, meus pensamentos levam-me, cada vez mais, na direção daquele rochedo tranquilo). A Figura 1.8 mostra a distribuição de frequências de alguns dados anuais inventados sobre o número de suicídios em Beachy Head, cometidos por pessoas de diferentes idades (embora eu tenha inventado esses dados, eles são baseados, grosseiramente, em estatísticas gerais de suicídio como os de Williams, 2001). Houve 171 suicídios no total e você pode ver que a população de suicidas tinha, mais frequentemente, idade entre 25

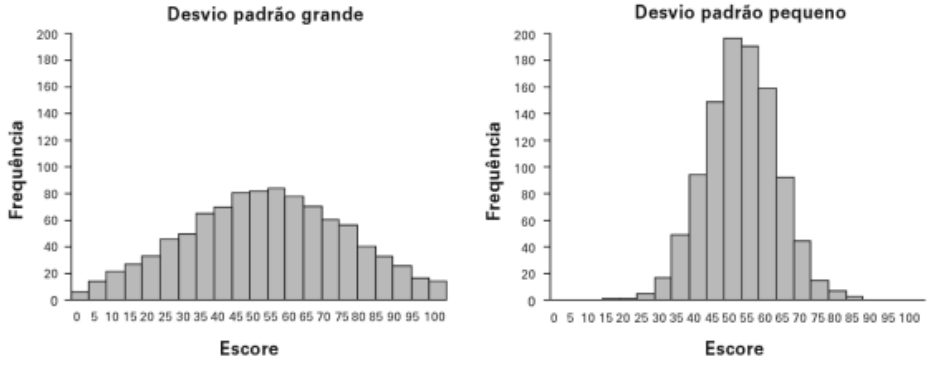


Figura 1.7 Duas distribuições com a mesma média, mas com desvios padrão grande e pequeno.

Material com direitos autorais

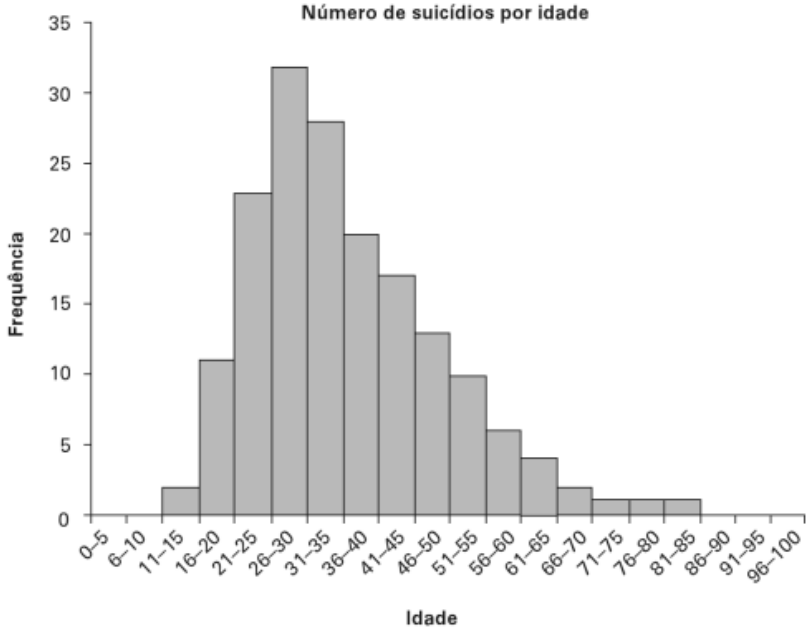


Figura 1.8 Distribuição de frequências mostrando o número de suicídios anuais por faixa etária.

e 30 anos (a barra mais alta), o que é bastante preocupante para mim (embora isso signifique que se eu conseguir passar por este ano a vontade irá diminuir no ano que vem!). O gráfico também nos diz que, por exemplo, poucas pessoas acima de 70 anos cometeram suicídio em Beachy Head.

Afirmei anteriormente que podíamos pensar nas distribuições de frequência em termos de distribuição de probabilidade. Para explicar isso, imagine que alguém pergunte a você o seguinte: "qual é a probabilidade de uma pessoa de 70 anos cometer suicídio em Beachy Head?". Qual seria a sua resposta? As chances são, se você olhar a distribuição de frequências, de sua resposta ser "pouco provável", pois há apenas duas pessoas do grupo de 171 suicidas com idade em torno de 70 anos. E se alguém perguntar "qual é a probabilidade de uma pessoa de 30 anos cometer suicídio?". De novo, olhando o gráfico, você poderá dizer "é bastante provável", pois 32 dos 171 suicidas eram pessoas com idade em torno de 30 anos (isso é mais do que uma em cada cinco pessoas que cometeram suicídio).

Assim, baseado na frequência de diferentes escores deve começar a ficar claro que podemos usar essa informação para estimar a probabilidade que um determinado escore irá ocorrer. Portanto, poderíamos perguntar, com base em nossos dados, qual é a probabilidade de uma vítima de suicídio ter entre 16 e 20 anos? O valor da probabilidade pode variar de 0 (não há chance alguma de o evento acontecer) a 1 (o evento definitivamente irá acontecer). Assim, por exemplo, quando eu falo com meus editores digo que há a probabilidade de 1 que irei completar a revisão deste livro por volta de junho de 2003. Entretanto, quando falo com outras pessoas, posso mais realisticamente dizer que há a probabilidade de 0,1 de eu terminar as revisões a tempo (ou seja, 10% de chance* ou 1 chance em 10 que eu completarei o livro no tempo previsto). Na verdade, a probabilidade de eu terminar no tempo

* N. de T.: Como você já deve ter notado, o autor utiliza o termo "chance" como sinônimo de probabilidade, o que nem todos o fazem e nem sempre é o caso.

Material com direitos autorais

previsto é 0 (nenhuma chance mesmo!) porque nunca consigo terminar na data prevista pelos editores! Se as probabilidades não fazem sentido para você, apenas ignore o ponto decimal e pense nelas como percentagens (isto é, a probabilidade de 0,1 de algo acontecer é o mesmo que 10% de chance de que algo aconteça).

Falei vagamente sobre como distribuições de frequências podem ser usadas para obter uma ideia aproximada da probabilidade de um escore ocorrer. Entretanto, é necessário ser preciso. Para qualquer distribuição de escores poderíamos, teoricamente, calcular a probabilidade de obter um escore de certo tamanho – seria muito entediante e complexo fazê-lo, mas poderíamos. Para poupar nossa sanidade, os estatísticos identificaram muitas distribuições. Para cada uma eles elaboraram uma fórmula matemática que especifica versões idealizadas dessas distribuições (elas estão especificadas em termos de uma curva). Essas distribuições idealizadas são conhecidas como distribuições de probabilidade e a partir delas é possível calcular a probabilidade de conseguir escores baseados nas frequências com que um escore particular ocorre. Umas dessas distribuições é a distribuição normal, já mencionada. Os estatísticos calcularam a probabilidade de certos escores ocorrerem numa distribuição normal com a média 0 e o desvio padrão de 1. Portanto, se tivermos qualquer conjunto de dados que possua a forma de uma distribuição normal e se a média e o desvio padrão são 0 e 1, respectivamente, podemos usar a tabela de probabilidades para a distribuição normal a fim de ver qual é a probabilidade de um determinado escore ocorrer nos dados (fiz uma tabela no Apêndice deste livro)*.

O problema óbvio é que nem todos os dados que coletamos terão uma média 0 e o desvio padrão 1! Por exemplo, podemos ter um conjunto que tem uma média de 567 e um desvio padrão de 52,98! Felizmente, qualquer conjunto de dados pode ser convertido em um conjunto que tenha a média 0 e o desvio padrão 1. Primeiro, para centrar os dados em zero, pegamos cada escore e subtraímos dele a média de todos os escores. Depois, dividimos o escore resultante pelo desvio padrão para assegurar que os resultados terão um desvio padrão de 1. Os escores resultantes são conhecidos como escores-z e, na forma de equação, a conversão que acabei de descrever é:

$$z = \frac{X - \bar{X}}{s}$$

A tabela das probabilidades que foi calculada para a distribuição normal padrão é mostrada no Apêndice. Por que essa tabela é importante? Bem, se olharmos os dados sobre suicídio, podemos responder a pergunta "qual é a probabilidade de alguém que se jogou de Beachy Head ter 70 anos ou mais?". Primeiro, convertemos 70 em um escore-z. Digamos que a média dos escores de suicídio foi 36 e o desvio padrão 13. O valor 70 irá tornar-se (70 – 36)/13 = 2,62. Depois, olhamos esse valor na coluna chamada "porções menores" (isto é, a área acima do valor 2,62). Devemos encontrar que a probabilidade é 0,0044, ou seja, somente 0,44% de chance que uma vítima de suicídio tenha 70 anos ou mais. Olhando a coluna chamada de "porções maiores", também podemos ver a probabilidade de que a vítima de suicídio tenha 70 anos ou menos! Essa probabilidade é 0,9956, ou seja, há 99,56% de chance de que uma vítima de suicídio tenha menos de 70 anos!

Espero que você possa ver a partir desses exemplos que a distribuição normal e os escores-z nos permitem ir, um passo além dos nossos dados, no sentido que de um conjunto de escores podemos calcular a probabilidade que um determinado escore irá acontecer. Assim, podemos ver se os escores de certo tamanho podem ou não ocorrer em uma distribuição de um tipo particular. Você verá quão útil isso é em seu devido tempo, mas vale mencionar, neste estágio, que certos escores-z são particularmente importantes. Isso porque seus valores são pontos de corte destacando percentagens importantes da distribuição. O primeiro

* N. de T.: O autor está falando de uma forma bastante informal, pois, de fato, a probabilidade de um valor específico ocorrer em uma distribuição normal é zero. A probabilidade só é diferente de zero em um intervalo de valores. Note que ele, ao calcular, utiliza intervalos (70 ou mais e 70 ou menos).

Material com direitos autorais

valor importante de z é 1,96, pois ele separa os 2,5% do topo da distribuição e o seu oposto (−1,96) destaca os 2,5% da cauda inferior da distribuição. Assim, tomados juntos, esses valores destacam 5% dos escores ou, colocando de outra maneira, 95% dos escores-z estão entre −1,96 e +1,96. Os outros dois importantes pontos de referência são os valores ±2,58 e ±3,29, que destacam 1% e 0,1% dos escores, respectivamente. Ou seja, 99% dos escores-z estão entre −2,58 e 2,58 e 99,9% deles estão entre −3,29 e 3,29. Lembre-se desses valores porque eles irão aparecer com frequência!

1.6 A MINHA AMOSTRA É REPRESENTATIVA DA POPULAÇÃO? ①

1.6.1 O erro padrão ①

Vimos que o desvio padrão e as distribuições das frequências nos dizem algo sobre quão bem a média representa aqueles dados, mas mencionei anteriormente que em geral coletamos dados de amostras em vez de toda uma população. Também falei que se você retirar várias amostras de uma população, elas irão diferir um pouco uma da outra. Portanto, é também importante saber quão bem uma amostra em particular representa a população. Aqui é onde usamos o erro padrão. Muitos estudantes não entendem a diferença entre o desvio padrão e o erro padrão (em geral porque essa diferença nunca é explicada claramente). Contudo, o erro padrão é um conceito importante que deve ser entendido, portanto, farei o possível para explicá-lo.

Já aprendemos que cientistas sociais usam amostras como uma forma de estimar o comportamento de uma população. Imagine que estamos interessados nos índices de todos os professores (assim, professores em geral é a população). Podemos pegar uma amostra dessa população. Quando alguém toma uma amostra da população, está pegando uma de muitas amostras possíveis. Se pegarmos muitas amostras de uma mesma população, cada amostra terá a sua própria média e em várias dessas amostras as médias serão diferentes (nem toda amostra terá a mesma média).

A Figura 1.9 ilustra o processo de selecionar amostras de uma população. Imagine que podemos ter os índices de todos os professores do planeta e que o índice médio é 3 (essa é a média da população). É claro que não podemos coletar os índices de todos os professores, portanto, usamos uma amostra. Para cada uma dessas amostras podemos calcular uma medida de posição ou média da amostra. Vamos imaginar que selecionamos nove amostras diferentes (como no diagrama). Você pode ver que algumas das amostras têm a mesma média que a população e algumas têm médias diferentes; a primeira amostra de professores teve média igual a 3, mas a segunda teve média igual a 2. Isso ilustra a variação amostral, isto é, as médias irão variar porque elas contêm elementos diferentes da população; a amostra que por acaso incluir alguns bons professores irá ter uma média maior do que uma amostra que, por acaso, incluir alguns péssimos professores! Podemos representar as médias amostrais como uma distribuição de frequências ou histograma[3], exatamente como fiz no diagrama.

Essa distribuição mostra que três amostras tinham a média 3, médias de 2 e 4 ocorreram em duas amostras cada e médias de 1 e 5 ocorreram apenas em uma amostra cada. O resultado final é uma bela distribuição simétrica conhecida como distribuição amostral. Uma distribuição amostral é simplesmente uma distribuição de frequências das médias de todas as amostras de uma mesma população. Teoricamente, selecionaríamos centenas ou milhares de amostras para construir uma distribuição amostral, mas estou usando apenas nove para simplificar o diagrama! A distribuição amostral apresenta o comportamento das amostras da população e você notará que ela está centrada

[3] Ele é um gráfico de cada média amostral representada contra o número de amostras que apresentam essa média – ver o capítulo 2 para detalhes.

Material com direitos autorais

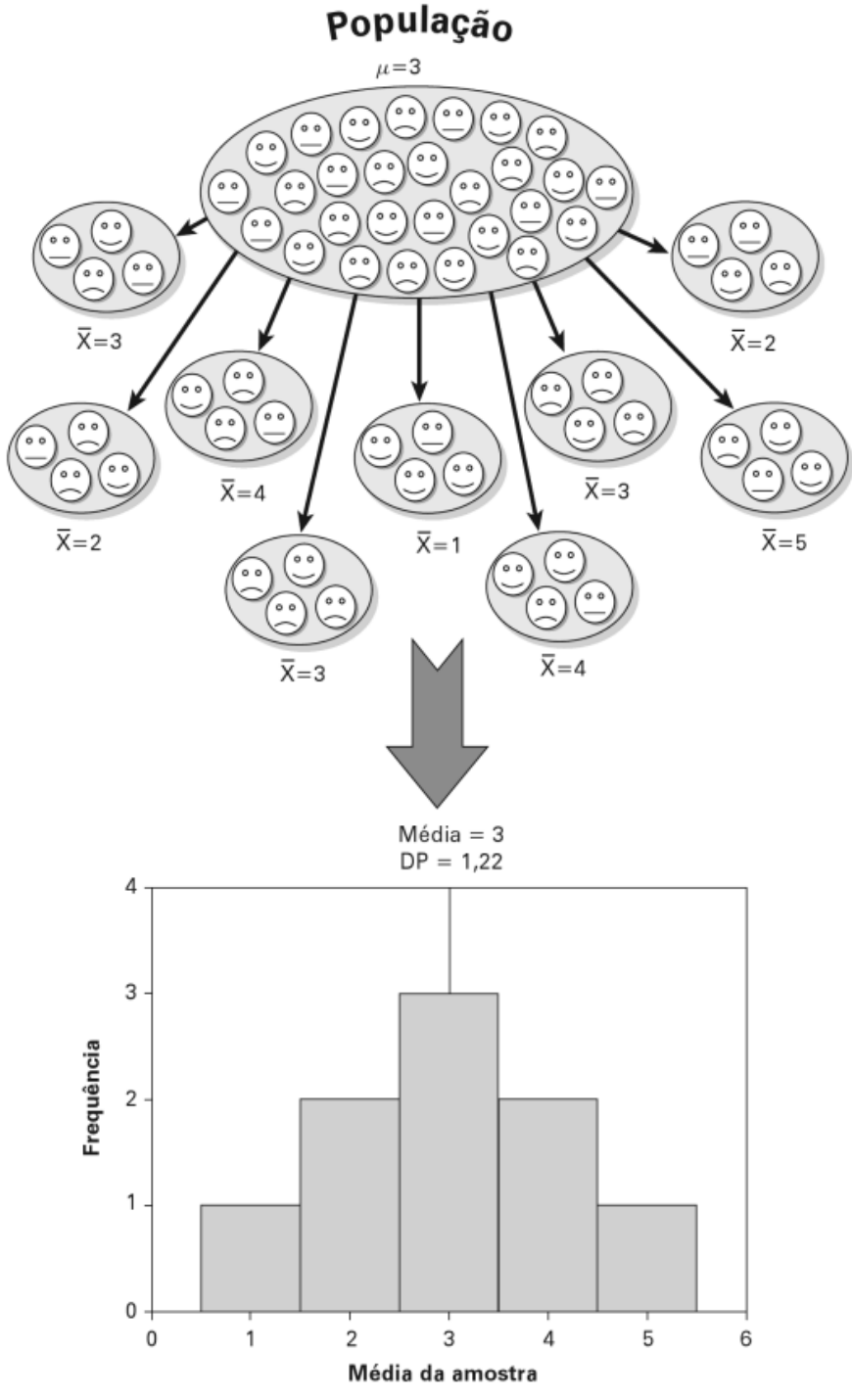


Figura 1.9 Figura ilustrando o erro padrão (veja o texto para detalhes).

Material com direitos autorais

no mesmo valor que a média da população (isto é, 3). Isso significa que, se calcularmos a média de todas as médias das amostras, teremos o mesmo valor da média da população. Agora, se a média das médias das amostras é igual à média da população e se conhecermos com precisão aquela média, saberemos algo sobre quão provável é que uma amostra qualquer represente a população. Então, como determinamos a precisão da média da população?

Lembre a discussão sobre o desvio padrão. Usamos o desvio padrão como uma medida de quão representativa a média é dos dados observados. Pequenos desvios padrões representam um cenário no qual a maioria dos dados está próxima da média e um desvio padrão grande representa uma situação na qual os dados estão bem mais espalhados em torno da média. Se você calcular o desvio padrão entre as médias das amostras, isso fornecerá uma medida de quanta variabilidade existe entre as médias de diferentes amostras. O desvio padrão entre as médias das amostras é denominado erro padrão da média (EP). Portanto, o erro padrão poderia ser calculado fazendo a diferença entre cada média da amostra e a média geral (média das médias de todas as amostras ou média da população), elevando ao quadrado essas diferenças, somando-as, e, então, dividindo-as pelo número de amostras.

É claro que, na realidade, não podemos selecionar centenas de amostras e, assim, nos baseamos numa aproximação do erro padrão (felizmente, muitos estatísticos espertos calcularam maneiras em que o erro padrão pode ser estimado a partir do desvio padrão da amostra). Felizmente, não precisamos entender por que essa aproximação funciona. Podemos apenas confiar que essas pessoas são inteligentes e sabem do que estão falando. O erro padrão pode ser calculado dividindo o desvio padrão da amostra (s) pela raiz quadrada do tamanho da amostra (N):[4]

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

1.6.2 Intervalos de confiança ②

Lembre que normalmente estamos interessados em utilizar a média da amostra como uma estimativa do valor da média verdadeira (isto é, da média da população). Acabamos de ver que amostras diferentes fornecerão valores diferentes da média e que podemos usar o erro padrão para ter uma ideia da extensão da diferença da média da amostra. Uma abordagem diferente para determinar a precisão da média da amostra como uma estimativa da média da população é calcular os limites entre os quais acreditamos que o valor da média verdadeira estará. Tais limites são chamados de intervalos de confiança. A ideia básica por trás dos intervalos de confiança é construir uma gama de valores dentro dos quais achamos que o valor da população estará.

[4] De fato, ele deve ser o desvio padrão da população (σ), que é dividido pela raiz quadrada do tamanho da amostra; no entanto, para amostras grandes essa é uma aproximação bastante boa.

Dica da Samanta Ferrinho

Resumindo, o erro padrão é o desvio padrão das médias das amostras. Como tal, ele é uma medida de quão representativa a amostra poderá ser da população. Um erro padrão grande (comparado à média da amostra) informa que existe muita variabilidade entre as médias das diferentes amostras e, dessa forma, a amostra que temos pode não ser representativa da população. Um erro padrão pequeno indica que muitas médias amostrais são similares (estão próximas) à média da população e, assim, a nossa amostra será provavelmente uma boa representação da população.

Material com direitos autorais

Vamos imaginar um exemplo: Domjan, Blesbois e Williams (1998) examinaram a liberação de esperma de uma codorna japonesa. A ideia básica é que se a codorna pode copular com a fêmea em um certo contexto (em uma câmara experimental), então esse contexto irá servir como uma deixa para a cópula e isso irá afetar a liberação do sêmen (embora durante a fase de testes a pobre codorna foi induzida a copular com um pedaço de pano com uma cabeça embalsamada de uma fêmea fincada no topo – doentio, não?). De qualquer modo, se olharmos a quantidade média de espermatozoides liberado na câmera experimental, existe uma média verdadeira (a média na população); vamos imaginar que são 15 milhões de espermatozoides. Agora, na nossa amostra real, poderemos constatar que a quantidade média de espermatozoide liberado foi de 17 milhões. Como não conhecemos a média verdadeira, realmente não sabemos se o valor da nossa amostra de 17 milhões é uma estimativa boa ou ruim desse valor. O que podemos fazer é usar um intervalo de estimativas: usamos nosso valor amostral como ponto do meio, mas determinamos um limite inferior e um superior. Assim, podemos dizer que o valor real da média de espermatozoide liberado é algo entre 12 e 22 milhões (note que os 17 milhões estão exatamente entre esses valores). É claro, nesse caso, o valor real (15 milhões) está entre esses limites. Mas e se determinarmos limites menores, por exemplo, se acharmos que os valores estão entre 16 e 18 milhões (novamente, note que os 17 milhões estão no meio)? Nesse caso, o intervalo não contém o valor real da média. Vamos imaginar que você está particularmente interessado no exemplo do espermatozoide da codorna japonesa e repetiu o experimento 50 vezes usando diferentes amostras. Cada vez que você realizou o experimento, construiu um intervalo em volta da média da amostra como acabei de descrever. A Figura 1.10 mostra esse cenário: os círculos representam a média para cada amostra com as linhas saindo de fora deles representando os intervalos para essas médias. O valor real da média (a média da população) é de 15 milhões e é representada pela linha vertical. O primeiro aspecto a ser notado é que a maioria das médias das amostras é diferente da média real (isso é por causa da variação da amostragem, como foi descrito na seção anterior). Em segundo lugar, embora a maioria dos intervalos contenha a média real (eles atravessam a linha vertical significando que 15 milhões de espermatozoides estão em algum lugar entre os limites mais altos e mais baixos), alguns não contêm.

Até agora, evitei o assunto de como devemos calcular os intervalos. É crucial construir os intervalos de confiança de uma maneira que eles nos informem algo útil. Portanto, nós os calculamos de uma forma que eles tenham certas propriedades: especificamente, que eles nos indiquem a probabilidade de conterem o valor real do que estamos tentando estimar (nesse caso, a média).

Tipicamente, se prestarmos atenção aos intervalos de confiança de 95% e, algumas vezes, aos intervalos de confiança de 99%, veremos que eles têm interpretações semelhantes: são limites construídos para que em certa percentagem das vezes (seja 95% ou 99%) o valor real da média da população esteja dentro desses limites. Assim, quando você tiver um intervalo de confiança de 95% para uma média, pense nele assim: se selecionarmos 100 amostras, calcularmos a média e, depois, determinarmos o intervalo de confiança para aquela média (parecido com a Figura 1.10), 95% dos intervalos de confiança conterão o valor real da média da população.

Para calcular o intervalo de confiança, precisamos saber os limites nos quais 95% das médias estarão. Como calculamos esses limites? Lembre-se da seção 1.5.3, onde eu disse que 1,96 era um valor de z importante (um escore de uma distribuição normal com uma média 0 e um desvio padrão 1) porque 95% dos escores-z estão entre $-1{,}96$ e 1,96. Isso significa que se a nossa média da amostra têm média 0 e erro padrão 1, os limites de nosso intervalo de confiança serão $-1{,}96$ e $+1{,}96$. Você deve também lembrar que podemos converter escores em escores-z usando esta equação:

$$z = \frac{X - \bar{X}}{s}$$

Material com direitos autorais

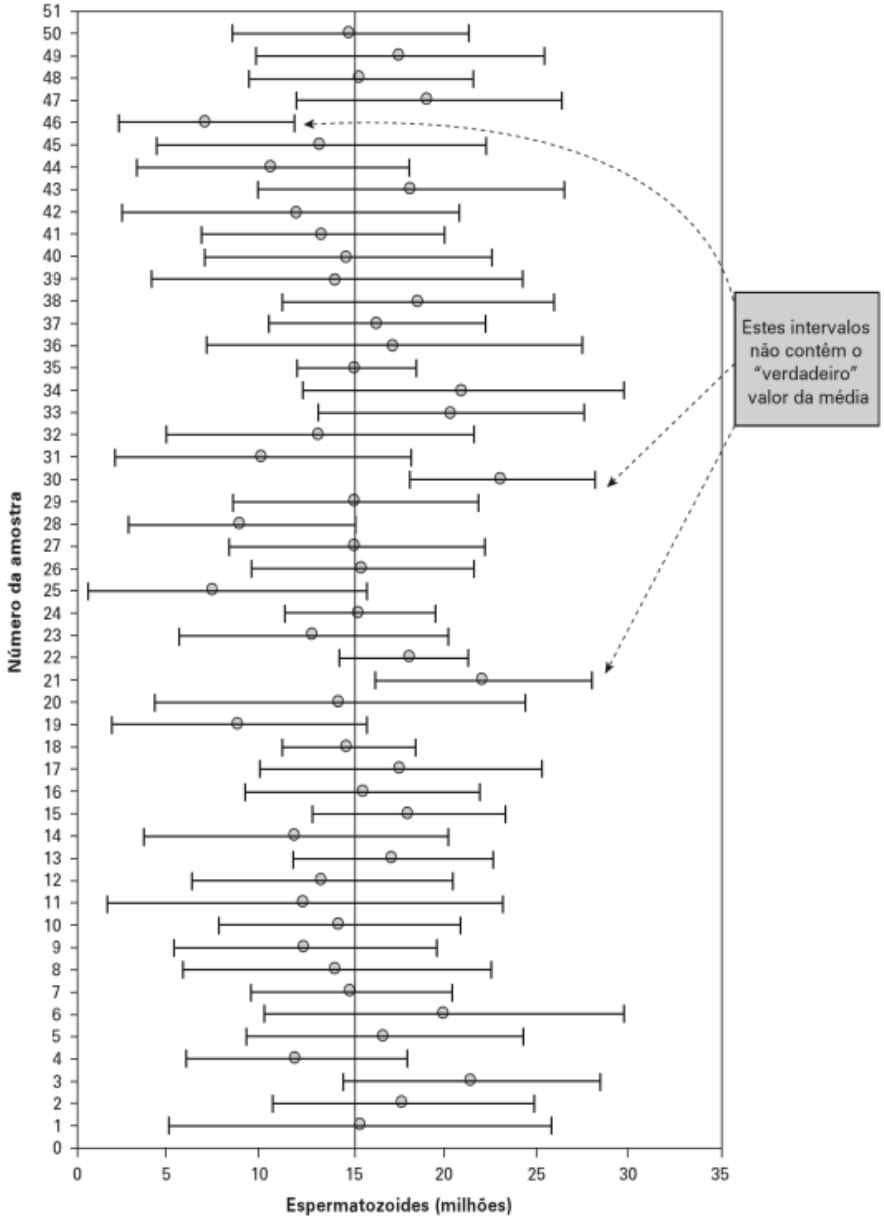


Figura 1.10 Figura mostrando intervalos de confiança da contagem de espermatozoides da codorna japonesa (eixo horizontal) para 50 diferentes amostras (eixo vertical).

Se soubermos que nossos limites serão −1,96 e 1,96 em escores-*z*, quais são os escores correspondentes em valores dos nossos dados? Para encontrar isso, podemos recolocar *z* na equação (porque existem dois valores, nós temos duas equações):

Material com direitos autorais

$$1{,}96 = \frac{X - \bar{X}}{s} \qquad -1{,}96 = \frac{X - \bar{X}}{s}$$

O que precisamos saber é o valor de X nessas equações e, para descobrir isso, nós simplesmente as reorganizamos:

$$1{,}96 \times s = X - \bar{X} \qquad -1{,}96 \times s = X - \bar{X}$$
$$1{,}96 \times s + \bar{X} = X \qquad -1{,}96 \times s + \bar{X} = X$$

Portanto, o intervalo de confiança pode ser facilmente calculado uma vez que o desvio padrão (*s* nas equações acima) e a média ($\bar{X}$ nas equações) são conhecidos. Na verdade, usamos o erro padrão e não o desvio padrão porque estamos interessados na variabilidade das médias das amostras e não na variabilidade das observações dentro da amostra. O limite mais baixo do intervalo de confiança é, portanto, a média menos 1,96 vezes o erro padrão e o limite superior é a média mais 1,96 erros padrão.

Limite inferior do intervalo de confiança
$= \bar{X} - (1{,}96 \text{ x EP})$

Limite superior do intervalo de confiança
$= \bar{X} + (1{,}96 \text{ x EP})$

Assim, a média está sempre no centro do intervalo de confiança. Se a média representa com precisão a média real, o intervalo de confiança deverá ser pequeno. Porque 95% dos intervalos de confiança contem a média real, podemos assumir que esse intervalo de confiança contém a média real: portanto, se o intervalo é pequeno, a média da amostra deve estar perto da média real. Já se o intervalo de confiança é muito grande, então a média da amostra deverá ser diferente da média real, indicando que ela é uma péssima representante da população. Você verá que os intervalos de confiança aparecerão frequentemente ao longo do livro.

1.7 MODELOS LINEARES ①

A média é um exemplo de um modelo estatístico, mas você pode perguntar que outros tipos de modelos estatísticos podem ser construídos. Bem, verdade seja dita, há somente um modelo geralmente usado, conhecido como modelo linear. Para alguns cientistas sociais, pode não ser óbvio que minha afirmação anterior está correta, mas um estatístico concordaria comigo prontamente. A razão para isso é que há uma variedade de nomes diferentes dados aos procedimentos estatísticos que são baseados no modelo linear. Um exemplo clássico é que a análise de variância (ANOVA) e a regressão são sistemas idênticos (Cohen, 1968), mas eles têm nomes diferentes e são usados em contextos diferentes (devido a uma divisão nas filosofias metodológicas – ver Cronbach, 1957).

A palavra linear literalmente significa "relativo a uma linha", mas em termos estatísticos, a linha referida é uma linha reta. Um modelo linear é, portanto, um modelo baseado sobre uma linha reta; isso significa que geralmente estamos tentando resumir nossos dados observados em temos de uma linha reta. Por exemplo, no capítulo descrevendo regressão, ficará claro que duas variáveis podem ser negativamente relacionadas (isso significa que à medida os valores de uma variável aumentam, valores de outra variável diminuem). Em tais circunstâncias, o relacionamento pode ser resumido por uma linha reta.

Suponha que nós mensuramos quantos capítulos deste livro uma pessoa leu e, então, medimos seu enriquecimento espiritual; podemos representar esses dados hipotéticos na forma de um diagrama de dispersão no qual cada ponto representa um escore individual em ambas variáveis. A Figura 1.11 mostra tal gráfico e também mostra o mesmo gráfico, mas com uma linha que resume o padrão desses dados. Uma terceira versão do diagrama de dispersão é também incluída, mas tem uma curva para resumir o padrão geral dos dados. A Figura 1.11 ilustra como podemos adequar diferentes tipos de modelos aos mesmos dados. Nesse caso podemos usar uma linha reta para

Material com direitos autorais

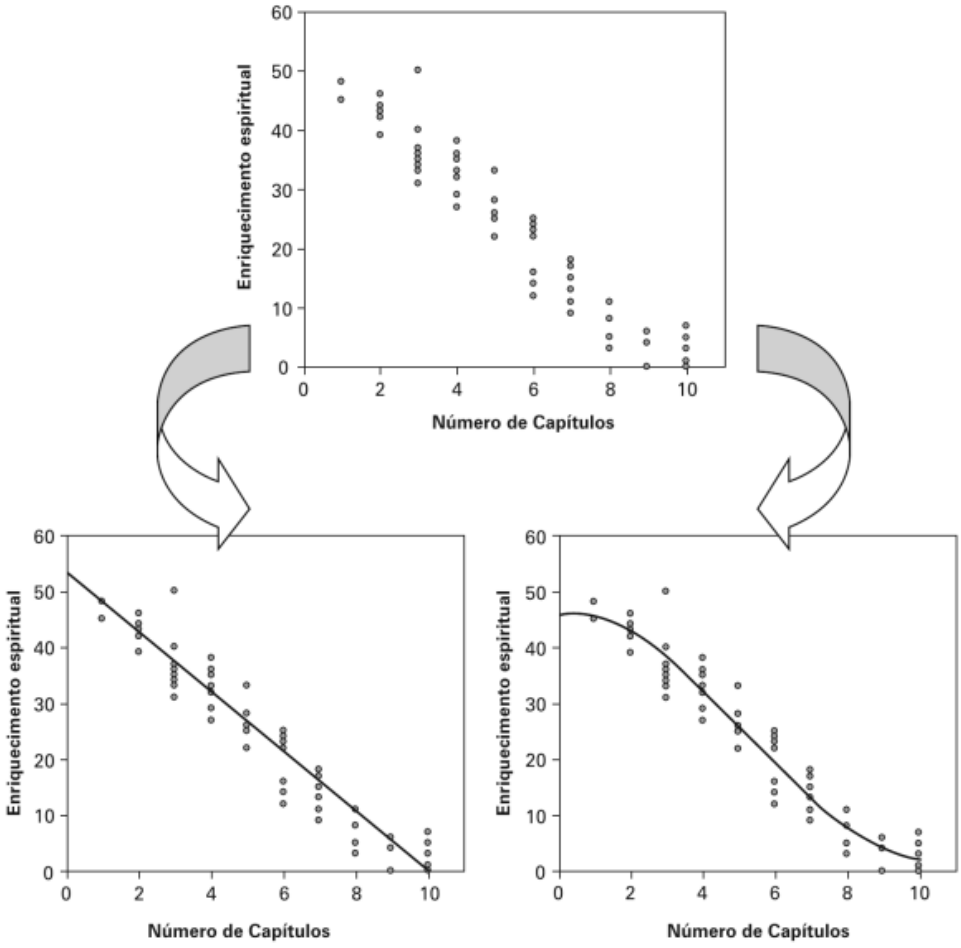


Figura 1.11 Diagramas de dispersão de alguns dados sem um modelo, com um modelo linear e um com um modelo não-linear ajustados.

representar nossos dados e ela irá mostrar que quanto mais capítulos uma pessoa ler, menor será o seu enriquecimento espiritual. Entretanto, também podemos usar uma linha curva para resumir os dados e isso mostrará que quando a maioria ou todos os capítulos foram lidos, o enriquecimento espiritual aumentará levemente (provavelmente porque uma vez que o livro for lido, tudo repentinamente faz sentido – imagine!). Nenhum dos dois tipos de modelos é necessariamente correto, mas existem casos em que um modelo representa melhor os dados do que outro. Assim, quando usamos modelos estatísticos é importante avaliar quão bem um determinado modelo adere aos dados.

A maioria da estatística usada nas ciências sociais é baseada em modelos lineares, o que significa que tentamos adequar modelos de linhas retas aos dados coletados. Isso é interessante porque a maioria dos estudos científicos publicados são aqueles com resultados estatísticos significantes. Dado que a maioria dos cientistas sociais aprende a usar somente técnicas baseadas no modelo linear, os resulta-

Material com direitos autorais

dos publicados serão aqueles que utilizarem os modelos lineares com sucesso. Dados que se ajustam a padrões não-lineares provavelmente serão erroneamente ignorados (porque o modelo errado teria sido aplicado aos dados, levando a resultados não-significantes). Por essa razão, bons pesquisadores primeiro representam seus dados graficamente; diagramas dos dados nos dizem muito sobre quais modelos podem ser aplicados aos dados. É possível, portanto, que algumas áreas da ciência estejam sendo tendenciosas; se você coletar dados que parecerem não-lineares, por que não corrigir a tendência e investigar técnicas estatísticas diferentes (o que é fácil para eu dizer quando não coloco tais técnicas no meu livro!)?

1.8 COMO DESCOBRIR SE O SEU MODELO REPRESENTA O MUNDO REAL? ①

Até agora, vimos que usamos amostras para estimar o que está acontecendo em uma população grande a qual não temos acesso. Vimos, também, que é importante estabelecer se um modelo é bom ou ruim para os dados e se ele é representativo da população. Expliquei tudo isso usando a média como um exemplo, mas o que acontece quando nosso modelo é mais complicado? Cientistas estão geralmente interessados em cenários mais complexos, como "existe uma relação entre o monte de bobagens que as pessoas falam e o monte de gelatina de vodka que elas comeram?" ou "a quantidade média de chocolate que eu como quando estou escrevendo o livro é maior do que quando não estou escrevendo?". Nesses casos, adequamos modelos que são mais complexos: na verdade, estamos detectando efeitos na nossa população e quantificando esses efeitos. Esta seção explica algumas maneiras de quantificar efeitos e decidir se eles são significativos. Essencialmente, esse é um processo de quatro estágios (veja o Quadro 1.1):

1. Crie uma hipótese (ou hipóteses) – isso será, geralmente, uma previsão de que algum tipo de efeito existe na população.
2. Colete alguns dados úteis.
3. Ajuste um modelo estatístico aos dados – esse modelo irá testar suas previsões originais.

Quadro 1.1

Fraude em pesquisa ①

O processo de quatro estágios descrito neste capítulo funcionará somente se você seguir os passos sem falcatruas. Suponha que eu quero apostar em quem vencerá a Copa do Mundo de Rúgbi. Como sou inglês, apostarei que a Inglaterra vencerá o torneio. Para fazer isso eu:

1. Faço minhas apostas, escolhendo meu time (Inglaterra) e as chances disponíveis na lotérica (atualmente 6/4).
2. Vejo qual time vence o torneio.
3. Coleto meus ganhos (se a Inglaterra vencer).

Para todos ficarem felizes, esse processo precisa ser justo: as lotéricas determinam suas chances de modo que não paguem muito (o que as torna felizes), mas às vezes paguem algo (para manter os clientes felizes). A lotérica pode oferecer qualquer chance antes de o torneio terminar, mas não pode mudá-las depois que o torneio acabar (ou o último jogo iniciou). Da mesma forma, posso escolher qualquer time antes do torneio, mas não posso mudar de ideia no meio do torneio ou no final do jogo!

(*Continua*)

Material com direitos autorais

Quadro 1.1 (*Continuação*)

A situação na pesquisa é semelhante: podemos escolher qualquer hipótese (time de rúgbi) que queremos antes de os dados serem coletados, mas não podemos mudar de ideia no meio da coleta dos dados (ou após a sua coleta). Da mesma forma, temos que decidir nosso nível de probabilidade (ou chances de apostas) antes de coletarmos os dados. Se fizermos isso, o processo de quatro estágios funciona. Entretanto, pesquisadores às vezes fraudam. Eles não escrevem suas hipóteses antes de conduzirem seus experimentos e, algumas vezes, eles as trocam quando os dados são coletados (como eu, trocando meu time depois que a Copa do Mundo terminou) ou, pior ainda, decidem sobre elas depois de coletar os dados! Com exceção de alguns procedimentos complicados chamados de testes *post hoc*, isso é fraude. Do mesmo modo, pesquisadores são culpados se escolherem o nível de significância depois de os dados terem sido coletados e analisados, assim como a lotérica que muda as chances depois do torneio.

Cada vez que você trocar sua hipótese ou os detalhes da sua análise, aumenta a chance de encontrar um resultado significativo, mas você também aumenta a probabilidade de publicar resultados que outro pesquisador não conseguirá reproduzir (o que é muito constrangedor). Se, entretanto, você seguir cuidadosamente as regras e fizer seu teste com uma significância de 5%, pelo menos saberá que no máximo um resultado em cada 20 poderá sair errado.

(Agradecimentos a David Hitchin por esse quadro – peço desculpas por torná-lo em um exemplo de rúgbi!).

4. Avalie esse modelo para ver se ele suporta suas previsões iniciais.

Cientistas, em geral, estão interessados em testar hipóteses; isto é, testar as questões científicas que eles criam. Dentro dessas questões há, usualmente, a previsão que um pesquisador fez. Essa previsão é chamada de hipótese experimental (ela é a previsão de que sua manipulação experimental terá algum efeito ou que de certas variáveis irão se relacionar entre si). A possibilidade contrária – que sua previsão está errada e que o efeito previsto não existe – é chamada de hipótese nula. Dois exemplos são:

1. *Hambúrgueres engordam*: a hipótese experimental é que quanto mais hambúrgueres você comer, mais parecido com uma baleia você vai ficar; a hipótese nula é que as pessoas irão ficar gordas independentemente da quantidade de hambúrgueres que comerem.
2. *Queijo dá pesadelos*: a hipótese experimental é que aqueles que comem queijo antes de dormir têm mais pesadelos do que quem não come; a hipótese nula seria que quem come queijo antes de dormir tem mais ou menos o mesmo número de pesadelos do que aqueles que não. (No caso de você estar interessado, eu frequentemente como queijo antes de dormir e não tenho pesadelos – exceto quando estou escrevendo livros de estatística e sonho que sou atormentado por um número mau.)

Uma grande parte deste livro aborda a estatística inferencial, que informa se a hipótese experimental pode ser verdadeira – ela nos ajuda a confirmar ou rejeitar nossas previsões. *Grosso modo*, aplicamos nosso modelo estatístico aos nossos dados e observamos quão bem ele se encaixa nos mesmos (em termos de quanta variância ele explica). Se ele se ajusta bem aos dados (isto é, explica muito da variância dos escores), assumimos que nossa previsão inicial é verdadeira: aceitamos a hipótese experimental. É claro, nunca podemos ter absoluta certeza de que qualquer hipótese está correta e, portanto, trabalhamos com probabilidades. Mais precisamente, calculamos a probabilidade de que os resultados que obtivemos ocorreram por acaso – à medida que essa probabilidade diminui, confirmamos que a hi-

Material com direitos autorais

pótese experimental é correta e que a hipótese nula pode ser rejeitada.

Para ilustrar essa ideia, Fisher[5](1925) descreve um experimento projetado para testar a afirmação de uma mulher de que ela conseguia determinar, provando uma xícara de chá, se o leite ou o chá foi colocado primeiro na xícara. Fisher pensou que poderia dar à mulher algumas xícaras de chá em que o leite fosse colocado primeiro e outras em que o leite fosse colocado por último e ver se ela conseguiria identificá-las corretamente. A mulher saberia que existiria um número igual de xícaras com o leite colocado em primeiro lugar e por último, mas não saberia a ordem em que foi colocado. Se pegarmos a situação mais simples na qual há somente duas xícaras, a mulher tem 50% de chance de adivinhar corretamente. Se ela realmente adivinhasse corretamente, não poderíamos ter certeza de que ela realmente sabe a diferença entre as xícaras nas quais o leite foi colocado primeiro daquelas em que o leite foi colocado por último, porque a maioria de nós poderia se sair bem nessa tarefa apenas adivinhando. Mas e se complicássemos as coisas colocando seis xícaras? Existem 20 maneiras nas quais essas xícaras podem ser arranjadas e a mulher somente adivinharia a ordem correta uma vez em 20 (ou 5% das vezes). Se ela acertasse a ordem correta, estaríamos provavelmente muito confiantes de que ela poderia, realmente, perceber a diferença (e reverenciá-la por seu apurado paladar). Se você quer saber mais sobre Fisher e suas palhaçadas de degustação de chá, veja Field e Hole (2003), mas, para os nossos propósitos, o ponto é que somente quando havia uma pequena probabilidade de a mulher completar a tarefa do chá somente por sorte concluiríamos que ela tinha uma habilidade genuína para detectar se o leite fora colocado na xícara antes ou depois do chá.

Não é coincidência eu ter escolhido o exemplo das seis xícaras acima (onde a provadora de chá tinha 5% de chance de acertar a tarefa por adivinhação), porque Fisher sugeriu que somente quando estamos 95% certos de que um resultado é genuíno (isto é, não resultante do acaso) devemos aceitá-lo como verdadeiro.[6] Ou seja, se há somente 5% de probabilidade de algo acontecer por acaso, podemos aceitar que é uma descoberta verdadeira – dizemos que é uma descoberta *estatisticamente significativa* (veja o Quadro 1.2 para entender melhor o que estatisticamente significativo quer dizer). Esse critério de 95% de confiança forma a base da moderna estatística, mas existe pouca justificativa para esse valor além do que Fisher disse; no entanto, como ele foi um sujeito muito inteligente, confiaremos no seu julgamento (Figura 1.12). Todavia, algumas vezes quando olho para a minha estante cheia de resultados experimentais que tiveram uma significância de somente 93% de probabilidade, imagino quão diferente minha carreira seria se Fisher tivesse acordado aquele dia com um humor tipo 90%.

Periódicos de pesquisas têm predisposição a publicar resultados positivos (nos quais a hi-

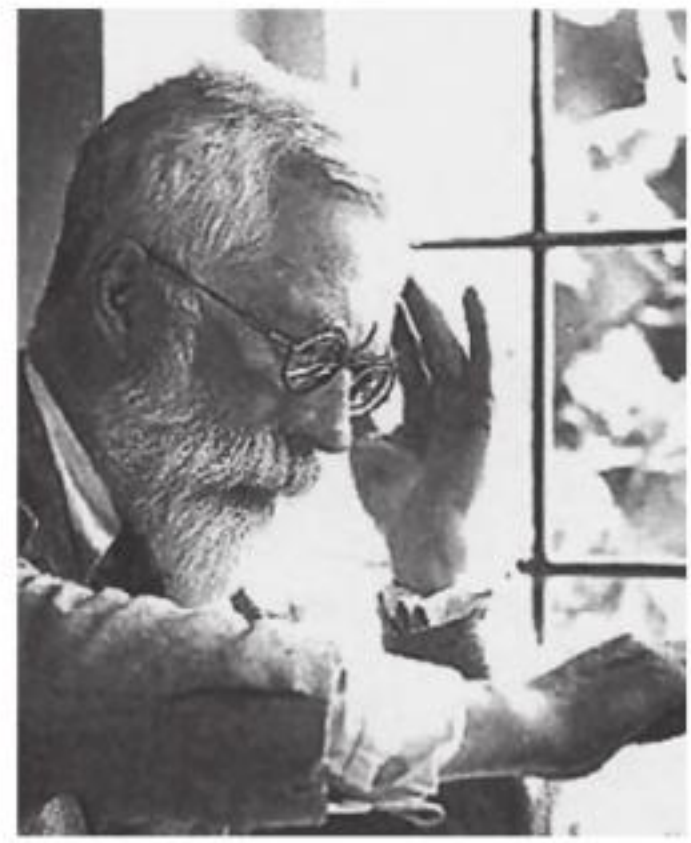


Figura 1.12 Ronald Fisher contemplando as consequências de estabelecer $p = 0,05$.

[5] Para saber mais sobre Fisher, visite o site http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafreader.htm.

[6] É claro, na realidade, ela pode não ser verdadeira – estamos apenas preparados para acreditar que ela seja!

Quadro 1.2

O que podemos ou não concluir de um teste de significância estatística ②

1. **A importância de um efeito:** Já vimos que a ideia básica por trás do teste de hipóteses envolve a geração de uma hipótese experimental e de uma hipótese nula, ajustar um modelo estatístico e avaliar aquele modelo com uma estatística teste. Se a probabilidade de obter o valor da nossa estatística teste por acaso for menor do que 0,05, então, geralmente aceitamos a hipótese experimental como verdadeira: há um efeito na população. Normalmente dizemos "existe um efeito *significativo* de...". Contudo, não seja enganado pela palavra "significativo", porque mesmo que a probabilidade do nosso efeito ter ocorrido por acaso seja pequena (menor do que 0,05), isso não quer dizer que o efeito é importante. Efeitos muito pequenos não-importantes podem se tornar estatisticamente significativos somente porque um grande número de pessoas foi usado no experimento (veja Field e Hole, p. 60).
2. **Resultados não-significativos:** Uma vez que você calculou sua estatística teste, você calcula a probabilidade de que ela tenha ocorrido por acaso; se essa probabilidade é maior do que 0,05 você rejeita sua hipótese experimental. Contudo, isso *não* significa que a hipótese nula é verdadeira. Lembre que a hipótese nula é que não existe efeito na população. Tudo o que um resultado não-significativo nos diz é que o efeito não é grande o suficiente para não ser outra coisa do que um resultado causal – ele não nos diz que o efeito é 0. Como Cohen (1990) destaca, um resultado não-significativo nunca deveria ser interpretado (apesar de sempre ser) como "nenhuma diferença entre médias" ou "nenhuma relação entre variáveis". Cohen também destaca que hipótese nula *nunca* é verdadeira porque sabemos pela distribuição de amostras (veja a Seção 1.6) que amostras ao acaso terão médias ligeiramente diferentes e, mesmo que essas diferenças sejam muito pequenas (isto é, uma média pode ser 10 e a outra 10.000001), elas são diferentes. Na verdade, mesmo uma pequena diferença seria considerada estatisticamente significante se uma amostra grande o suficiente fosse utilizada. Assim, testes de significância nunca nos dizem que a hipótese nula é verdadeira, porque ela nunca é!
3. **Resultados significativos:** OK, podemos não aceitar a hipótese nula como sendo verdadeira, mas podemos pelo menos concluir que ela é falsa quando nossos resultados são significativos, certo? Errado! Uma estatística teste significativa é baseada em uma argumentação probabilística, com grandes limites do que podemos concluir. Novamente, Cohen (1994), que era um ótimo escritor de estatística, destaca que uma argumentação formal se baseia em uma afirmação inicial de um fato seguida por uma afirmação sobre o atual estado dos assuntos e uma conclusão inferida. Este silogismo ilustra o que eu quero dizer:

   - Se um homem não tem braços ele não pode tocar guitarra.
     Este homem toca guitarra.
     Portanto, este homem tem braços.

O silogismo inicia com uma afirmação do fato que permite alcançar a conclusão final, porque você pode negar que o homem não tem braços (o anterior), negando que ele não pode tocar guitarra (o seguinte). Uma versão parecida de hipótese nula é:

- Se a hipótese nula é correta, esta estatística teste não pode acontecer.
  Esta estatística teste aconteceu.
  Portanto, a hipótese nula é falsa.

Isso é ótimo, mas a hipótese nula não é representada dessa maneira porque ela é baseada em probabilidades. Ela deveria ser colocada da seguinte forma:

- Se a hipótese nula está correta, esta estatística teste é pouco provável.
  Esta estatística teste aconteceu.
  Portanto, a hipótese nula é pouco provável.

(*Continua*)

Material com direitos autorais

Quadro 1.2 (*Continuação*)

Se voltarmos ao exemplo da guitarra, podemos ter uma afirmação similar:

- Se um homem toca guitarra, provavelmente ele não toca na banda Fugazi (isso é verdade porque há milhares de pessoas que tocam guitarra, mas somente duas que tocam guitarra na banda Fugazi!).
  Guy Picciotto toca na Fugazi.
  Portanto, Guy Piciotto provavelmente não toca guitarra.

Isso deve parecer, espero, completamente ridículo – a conclusão está errada porque Guy Piciotto toca guitarra. Isso ilustra um erro comum em testes de hipóteses. De fato, testes de significância nos permitem dizer pouco sobre a hipótese nula.

pótese experimental é apoiada) e, se Fisher tivesse acordado com um humor em 90% aquela manhã, eu teria tido muito mais experimentos "com sucesso" e provavelmente seria vice-reitor da minha universidade agora (hum, talvez tenha sido melhor para minha universidade que ele tenha acordado com um humor de 95%!).

1.8.1 Estatísticas teste ①

Então, como sabemos que o nosso modelo é uma boa representação do que está acontecendo no mundo real? Quando selecionamos dados, esses dados irão variar (vimos como medir essa variância na Seção 1.4.1) Há dois tipos de variância (veja também o Capítulo 7):

- **Variação sistemática:** Essa variação se deve a um efeito genuíno (seja um efeito de um pesquisador fazendo algo a todos os participantes de uma amostra, mas não em outras, ou a variação natural entre conjuntos de variáveis). Você pode pensar nisso como uma variação que pode ser explicada pelo modelo que ajustamos aos dados.
- **Variação não-sistemática**: É uma variação que não se deve ao efeito em que estamos interessados (assim, pode ser devido a diferenças naturais entre pessoas em diferentes amostras, como diferenças de inteligência e motivação). Você pode pensar nisso como uma variação que não pode ser explicada pelo modelo que ajustamos aos dados.

Se estivermos tentando estabelecer se um modelo é uma representação razoável do que está acontecendo na população, geralmente calculamos uma estatística teste. Uma estatística teste é uma estatística que tem propriedades conhecidas; nós sabemos a freqüência com que diferentes valores dessa estatística ocorrem. Sabendo isso, podemos calcular a probabilidade de obter um determinado valor. Uma analogia usada por Field e Hole (2003) é a idade em que as pessoas morrem. Dados passados nos dizem a distribuição da idade de morte. Por exemplo, sabemos que em média os homens morrem por volta dos 75 anos e essa distribuição é bem pesada no topo; isto é, a maioria das pessoas morre com idade acima dos 50 anos e é muito incomum morrer por volta dos 20. Assim, frequências da idade do falecimento em mais velhos são muito altas, mas mais baixas em mais jovens. Desses dados seria possível calcular a probabilidade de alguém morrer com certa idade. Se aleatoriamente pegarmos alguém e perguntarmos sua idade, e for 53, poderíamos dizer qual a probabilidade de ele morrer antes do seu próximo aniversário (mas provavelmente levaríamos um soco!). Também, digamos que encontramos um homem com 110 anos; poderíamos calcular qual é a probabilidade de ele ter vivido tanto (seria uma probabilidade muito pequena porque a maioria das pessoas morre antes de alcançar essa idade). A maneira como usamos estatísticas teste é muito similar: sabemos suas distribuições e isso nos permite, uma vez calculada a estatística teste, descobrir a probabilidade de achar um valor tão grande como o que temos. Assim, se calculássemos a estatística teste e seu

Material com direitos autorais

valor fosse 110, poderíamos, então, calcular a probabilidade de obter um valor tão grande.

Assim, como calculamos essas estatísticas testes? Isso depende de qual estatística você está usando (e como você verá mais adiante neste livro, existem muitos: t, F e χ^2, para listar apenas três). Entretanto, a maioria dessas estatísticas representa essencialmente a mesma coisa:

$$\text{Estatística teste} = \frac{\text{Variância explicada pelo modelo}}{\text{Variância não-explicada pelo modelo}}$$

A forma exata dessa equação muda de teste para teste, mas essencialmente estamos comparando a quantidade da variância explicada pelo modelo que adequamos aos dados à variância que não pode ser explicada pelo modelo (veja os Capítulos 5 e 7 para uma explicação mais detalhada). A razão por que essa proporção é tão útil é intuitiva: se nosso modelo é bom, esperamos que o número de variâncias que ele pode explicar seja maior do que o número que ele não pode explicar. Nesse caso, a estatística teste será sempre maior do que 1 (mas não necessariamente significante).

Dado que sabemos quão frequentemente diferentes valores dos testes estatísticos ocorrem por acaso, uma vez calculada uma estatística teste particular, podemos calcular a probabilidade de conseguir esse valor. Como vimos na seção anterior, esses valores de probabilidade nos dizem quão provável é que nosso modelo, ou efeito, seja genuíno e não somente resultado do acaso. Quanto mais variação nosso modelo explicar (comparado com a variância que ele não pode explicar), maior será o teste estatístico e menor a possibilidade de ele ocorrer por acaso (como nosso homem de 110 anos). Portanto, quanto maior a estatística teste, menor a probabilidade de eles ocorrerem. Quando essa probabilidade cai para abaixo de 0,05 (Critério de Fisher), aceitamos isso como uma confiança suficiente para assumir que a estatística teste é assim grande porque nosso modelo explica um montante suficiente de variações para refletir o que realmente está acontecendo no mundo real (a população). Em outras palavras, aceitamos nossa hipótese experimental e rejeitamos nossa hipótese nula – contudo, o Quadro 1.2 explica alguns mal-entendidos comuns sobre este processo.

1.8.2 Testes uni e bilaterais ①

Quando usamos testes estatísticos ou temos uma previsão específica do que irá acontecer, como "quanto mais alguém ler este livro maior a sua vontade de matar o autor", ou realmente não sabemos o que irá acontecer, como "ler mais deste livro poderá aumentar ou diminuir o desejo do leitor de matar o autor". O primeiro exemplo é direcionado: explicitamente dissemos que as pessoas iriam querer me matar à medida que lessem o livro. Se testarmos essa hipótese estatisticamente, o teste é chamado de **teste unilateral**. A segunda hipótese não é direcionada: afirmamos que o desejo de me matar irá mudar por causa da leitura do livro, mas não dissemos se esse desejo irá aumentar ou diminuir. Se testarmos essa hipótese estatisticamente, esse teste será **bilateral**.

Suponha que queremos descobrir se ler este livro aumenta ou diminui o desejo de me matar. Poderíamos fazer isso (experimentalmente) pegando dois grupos, um que leu o livro e outro que não leu, ou (correlacionalmente) medindo o montante do livro que foi lido e o desejo correspondente de me matar. Se não tivermos uma hipótese direcional, há três possibilidades. (1) As pessoas que leram este livro querem me matar mais do que as que não o leram, assim, a diferença (a média dos que leram o livro menos a média dos que não leram) é positiva. Em correlação, quanto mais você ler o livro, mais você irá querer me matar – uma relação positiva. (2) As pessoas que leram este livro querem me matar menos do que as que não leram, assim, a diferença (a média das pessoas que estão lendo o livro menos a das que não leram) é negativa. Em correlação, quanto mais você ler o livro, menor o seu desejo de me matar – uma relação negativa. (3) Não há diferença no desejo de me matar entre leitores e não-leitores – a média dos leitores menos a média dos não leitores é exatamente zero. Em correlação, não há relação entre ler este livro e a vontade de me

Material com direitos autorais

matar. Essa última opção é uma hipótese nula. A direção da estatística teste (isto é, se é positiva ou negativa) depende se a diferença é positiva ou negativa. Assumindo que há uma diferença positiva ou um relacionamento (ler este livro faz com que você queira me matar), então, para detectar essa diferença temos que levar em conta o fato de que a média para leitores é maior do que para não-leitores (e assim obtemos uma estatística teste positiva). Entretanto, se fizemos uma previsão incorreta e ler este livro faz com que leitores queiram me matar menos, a estatística teste será realmente negativa.

Quais as conseqüências disso? Bem, se no nível 0,05 precisássemos ter uma estatística teste maior do que, por exemplo, 10, e o que realmente temos é −12, então rejeitaríamos a hipótese mesmo que uma diferença exista. Para evitar isso, podemos olhar os dois lados (ou caudas) da distribuição de possíveis estatísticas teste. Isso significa que iremos pegar ambas as estatísticas teste, negativas ou positivas. Entretanto, fazer isso tem um preço, pois, para manter nosso critério de probabilidade em 0,05, devemos dividir essa probabilidade entre as duas caudas: assim, temos 0,025 no lado positivo da distribuição e 0,025 no lado negativo. A Figura 1.13 mostra essa situação – as áreas escurecidas são as áreas acima do necessário para o teste estatístico ao nível 0,025 de significância. Combine as probabilidades em ambos os lados e você terá 0,05 – nosso critério de valor. Agora, se fizemos uma previsão, colocamos todos nossos ovos em uma cesta e olhamos somente para o final da distribuição (ou para o lado positivo ou o negativo, dependendo na direção da previsão que fizemos). Consequentemente, podemos apenas procurar pelo valor de uma estatística teste que ocorreria por acaso com a uma probalidade de 0,05. Na Figura 1.13, as linhas diagonais mostram a área acima da estatística teste positiva necessária ao nível 0,05 de significância. O ponto importante é que se fizermos uma previsão específica, precisamos de uma estatística teste menor para encontrar um resultado significativo (porque estamos procurando somente em uma cauda), mas se nossa previsão está na direção errada, não conseguiremos detectar o efeito que realmente existe! Nesse contexto, é importante lembrar o que foi dito no Quadro 1.1: você não pode apostar ou mudar sua aposta depois que o torneio terminar – se você não fez a previsão da direção antes de coletar os dados, será tarde demais para prever a direção e reivindicar as vantagens do teste unilateral.

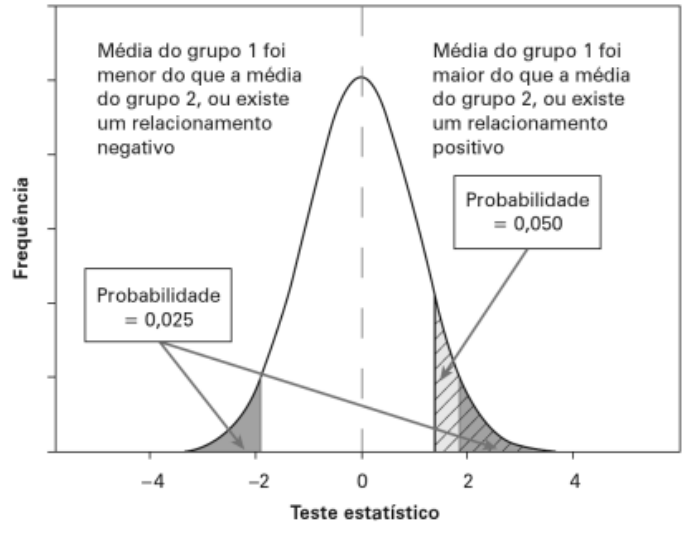


Figura 1.13 Diagrama para mostrar a diferença entre testes uni e bilaterais.

Material com direitos autorais

1.8.3 Erros do tipo I e do tipo II ①

Usamos testes estatísticos para nos informar sobre o estado real do mundo (a certo grau de confiança). Especificamente, estamos tentando ver se existe um efeito na nossa população. Temos duas possibilidades no mundo real: existe, na realidade, um efeito na população, ou não existe, na realidade, efeito na população. Não temos como saber quais dessas possibilidades são verdadeiras: entretanto, podemos olhar as estatísticas teste e suas probalidades associadas para descobrir qual dos dois é mais provável. Obviamente, é importante que sejamos o mais preciso possível, por isso Fisher afirmou que devemos ser muito conservadores e somente acreditar que o resultado é verdadeiro quando estivermos 95% confiantes de que ele é – ou quando houver somente 5% de chance de que os resultados possam ocorrer por acaso. Existem dois erros que podemos cometer: erro do Tipo I e erro do Tipo II. O **erro do Tipo I** ocorre quando acreditamos que há um efeito verdadeiro na nossa população e, de fato, não há. Se usarmos o critério de Fisher, a probabilidade de erro é 0,05 (ou 5%) quando não existe efeito na população – esse valor é conhecido como nível de significância e sua probabilidade é representada por α. Assumindo que não há efeito na população, e se reproduzirmos nossa coleção de dados 100 vezes, poderíamos esperar que em cinco ocasiões obtivéssemos uma estatística teste grande o suficiente para nos fazer pensar que há um efeito genuíno na população mesmo que ele não exista. O oposto é o erro do Tipo II, que ocorre quando acreditamos que não exista um efeito na população, mas na realidade, ele existe. Isso irá ocorrer quando obtivermos um valor da estatística teste pequeno (talvez porque exista muita variação natural entre nossas amostras). Em um mundo ideal, queremos que a probabilidade desse erro seja bem pequena (se existe um efeito na população então é importante que possamos detectá-lo). Cohen (1992) sugere que a probabilidade máxima aceitável para um erro do Tipo II seria 0,2 (ou 20%) – isso é denominado de nível-β. Isso significa que, se tomássemos 100 amostras de dados da população na qual um efeito existe, falharíamos em detectar esse efeito em 20 dessas amostras (assim, perderíamos 1 em 5 dos efeitos verdadeiros).

Existe, obviamente, uma troca entre esses dois erros: se diminuirmos a probabilidade de aceitar um efeito como verdadeiro (isto é, tornar α menor), aumentamos a probabilidade de que rejeitaremos um efeito que realmente existe (porque somos muito rígidos sobre o nível no qual aceitaremos um efeito como verdadeiro). A relação exata entre o erro do Tipo I e o erro do Tipo II não é direta porque eles estão baseados em diferentes suposições: para fazer um erro do Tipo I não deve haver efeito na população, enquanto para fazer um **erro do Tipo II** vale o contrário (deve haver um efeito que perdemos). Assim, embora saibamos que a probabilidade de cometer um erro do Tipo I diminui à medida que a probabilidade de cometer um erro do Tipo II aumenta, a natureza exata do relacionamento é geralmente adivinhada pelo pesquisador com base em sua experiência e conhecimento (Howell, 2002, p. 104-107, fornece uma ótima explicação sobre a troca entre erros).

1.8.4 Tamanhos de efeito ②

A estrutura para testar se os efeitos são verdadeiros que acabamos de apresentar apresenta alguns problemas – a maioria deles foi explicada brevemente no Quadro 1.2. O primeiro problema que encontramos foi saber quão importante um efeito é: apenas porque uma estatística teste é significativa não quer dizer que o efeito que ela mede é significativo ou importante. A solução é medir o tamanho de efeito que estamos testando de uma maneira padronizada. Quando medimos o tamanho de um efeito (seja uma manipulação experimental ou a força do relacionamento entre variáveis), isso é conhecido como tamanho de efeito. O tamanho de efeito é simplesmente uma medida de magnitude padronizada do efeito observado. O fato de que a medida é padronizada apenas significa que

Material com direitos autorais

podemos comparar os tamanhos dos efeitos por meio de diferentes estudos que mediram diferentes variáveis ou usaram medidas de escala distintas (assim, um tamanho de efeito baseado em velocidade de milésimos de segundos poderia ser comparado a um tamanho de efeito baseado em batimentos cardíacos). Muitas medidas do tamanho de efeito foram propostas, as mais comuns são o *d* de Cohen e o coeficiente de correlação *r* de Pearson (veja Field, 2001 e Capítulo 4). Muitos de vocês estão familiarizados com o coeficiente de correlação como uma medida da força do relacionamento entre duas variáveis (veja o Capítulo 4); contudo, ela é também uma medida muito versátil da força de um efeito experimental. É um pouco difícil aceitar que o humilde coeficiente de correlação também pode ser usado dessa maneira; entretanto, isso é somente porque os estudantes aprendem sobre ele dentro de um contexto de pesquisa não-experimental. Não quero falar sobre isso agora, mas se você ler os Capítulos 4, 7 e 8 entenderá melhor (eu espero!) o que quero dizer. Pessoalmente, prefiro o coeficiente *r* de correlação de Pearson como uma medida do tamanho de efeito porque ele está limitado ao intervalo entre 0 (sem efeito) e 1 (um efeito perfeito).[7]

Tamanhos de efeito são úteis porque eles dão uma medida objetiva da importância de um efeito. Assim, não importa qual é o efeito que você está procurando, quais as variáveis que foram medidas: sabemos que um coeficiente de correlação 0 significa que não existe efeito e um valor 1 significa que existe um efeito perfeito. Cohen (1988, 1992) estipulou o que é um efeito pequeno ou grande:

- $r = 0{,}10$ **(efeito pequeno):** nesse caso, o efeito explica 1% da variância total.
- $r = 0{,}30$ **(efeito médio):** o efeito é responsável por 9% da variância total.
- $r = 0{,}50$ **(efeito grande):** o efeito é responsável por 25% da variância total.

Podemos usar essas diretrizes para acessar a importância do nosso efeito (independentemente da significância da estatística teste). Entretanto, *r* não é medido numa escala linear, assim, um efeito com $r = 0{,}6$ não é duas vezes maior do que um com $r = 0{,}3$! A utilidade da estimativa do tamanho de efeito é tamanha que a *American Psychological Association* (APA) recomenda que todos os psicólogos informem esses efeitos nos resultados de qualquer trabalho publicado. Portanto, é um hábito que vale a pena ter.

Um ponto final a mencionar é que quando calculamos os tamanhos de efeito, nós os calculamos para uma determinada amostra. Quando olhamos as médias em uma amostra, vemos que as usamos para fazer inferências sobre a média de uma população inteira (que é o valor que realmente estamos interessados). O mesmo é verdadeiro sobre os tamanhos de efeito: o tamanho de efeito na população é o valor no qual estamos interessados, mas porque não temos acesso a esse valor, usamos o tamanho de efeito na amostra para estimar o provável tamanho de efeito na população (veja Field, 2001).

1.8.5 Poder estatístico ②

Vimos que os tamanhos de efeito são uma maneira inestimável de expressar a importância de uma descoberta na pesquisa. O tamanho de efeito em uma população é intrinsecamente ligado a três outras propriedades estatísticas: (1) o tamanho da amostra no qual o tamanho de efeito da amostra é baseado; (2) o nível da probabilidade no qual aceitaremos que um efeito é estatisticamente significativo (nível α); (3) a habilidade de um teste detectar um efeito daquele tamanho (conhecido como o poder estatístico). Assim, se conhecemos três dessas propriedades, podemos calcular a que falta. Irá depender, também, se o teste é uni ou bilateral (veja a Seção 1.8.2). Tipicamente, em psicologia, usamos o nível α em 0,05, assim, já conhecemos esse valor. O poder de um teste é a probabilidade que um determinado teste irá

[7] O coeficiente de correlação pode ter valores negativos (mas não abaixo de −1), o que é útil quando estamos medindo o relacionamento entre duas variáveis porque o sinal *r* nos diz sobre a direção do relacionamento, mas em pesquisas experimentais o sinal *r* meramente reflete a maneira como os pesquisadores codificam seus grupos (veja o Capítulo 4).

Material com direitos autorais

encontrar um efeito assumindo que um já exista na população. Se você refletir, irá lembrar que já vimos a probabilidade de falhar em detectar um efeito que realmente existe (β, a probabilidade de um erro Tipo II). Segue que a probabilidade de detectar um efeito, se um existe, deve ser o oposto da probabilidade de não detectar aquele efeito (isto é, $1 - \beta$). Também já mencionei que Cohen (1988, 1992) sugere que deveríamos esperar uma probabilidade 0,2 na falha em detectar um efeito verdadeiro e, assim, o nível correspondente do poder que ele recomendou foi de 1,0 – 0,2 = 0,8. Nosso objetivo deve ser alcançar um poder de 0,8 ou 80% de probabilidade de detectar um efeito se ele genuinamente existe. O tamanho de efeito na população pode ser estimado do tamanho de efeito em uma amostra e o tamanho da amostra é determinado pelo pesquisador de qualquer modo, assim, o valor é fácil de calcular. Agora, existem duas coisas úteis que podemos fazer sabendo que essas quatro variáveis estão relacionadas:

1. **Calcular o poder de um teste:** Se conduzirmos nosso experimento, já teremos selecionado um valor de α, poderemos estimar o tamanho de efeito baseado na nossa amostra e saberemos quantos participantes usamos. Portanto, podemos usar esses valores para calcular β, o poder de nosso teste. Se esse valor for 0,8 ou mais, podemos garantir que alcançamos poder suficiente para detectar qualquer efeito que poderá ter existido, mas se o valor resultante for menor, deveremos refazer o experimento usando mais participantes para aumentar o poder.
2. **Calcular o tamanho da amostra necessário para alcançar um nível dado de poder:** Dado que sabemos o valor de α e β, podemos utilizar pesquisas anteriores para estimar o tamanho de efeito que esperamos detectar em um experimento. Mesmo que ninguém tenha previamente feito o experimento exato que você pretende fazer, ainda assim podemos estimar a provável tamanho de efeito baseados em experimentos similares. Podemos usar esse tamanho de efeito estimado para calcular quantos participantes necessitamos para detectar esse efeito (baseado nos valores de α e β que escolhemos).

O último uso é o mais comum: para determinar quantos participantes devem ser usados para alcançar o desejado nível de poder. Os cálculos exatos são muito volumosos, mas, felizmente, existem agora programas de computadores disponíveis que os farão para você (um exemplo é o G* Power que é grátis e pode ser encontrado no *site* www.artmed.com.br, outro é o nQuery Adviser – veja Field (1998b) para uma avaliação – mas esse tem que ser comprado!) Também, Cohen (1988) fornece tabelas extensas para calcular o número de participantes para um determinado nível de poder (e vice-versa). Baseado em Cohen (1992), podemos usar as seguintes diretrizes: se tomarmos o nível padrão α de 0,05 e requisitarmos o poder recomendado de 0,8, então precisamos de 783 participantes para detectar um tamanho de efeito pequeno ($r = 0,1$), 85 participantes para detectar um tamanho de efeito médio ($r = 0,3$) e 28 participantes para detectar um tamanho de efeito grande ($r = 0,5$).

1.9 ÚLTIMOS CONSELHOS

Quero terminar este capítulo com alguns conselhos gerais. Procedimentos estatísticos são uma forma de processar números e, portanto, se você colocar baboseiras em uma análise mesmo assim obterá conclusões que tem significado estatístico, mas que provavelmente não terão sentido empírico (prático). Existe uma tentação de ver a estatística como uma forma milagrosa de determinar a verdade, mas a estatística é somente um recurso. Meu professor de estatística costumava dizer que se "você entra com lixo, você vai obter lixo". Eu nunca havia entendido o que ele estava querendo dizer até que as pessoas começaram a me apresentar conjuntos de dados contendo milhões de variáveis aleatórias aparentemente conectadas que eles queriam que eu as ajudasse a interpretar. Na análise estatística não existe um substituto para o pensamento empírico! Tenha isso em mente ao

Material com direitos autorais

longo deste livro. Um último conselho: nunca acredite em algo que um psicólogo diga sobre estatística. ☺

1.10 O QUE DESCOBRIMOS SOBRE ESTATÍSTICA? ①

Ok, este foi o seu curso rápido de iniciação à teoria da estatística! Espero que seu cérebro ainda esteja relativamente intacto. O ponto principal que quero que você entenda é que quando executa a sua pesquisa, você estará querendo ver se algum efeito verdadeiro existe na população (o efeito procurado depende dos seus interesses de pesquisa e suas previsões específicas). Você não será capaz de coletar dados sobre toda a população (a menos que queira gastar toda a sua vida e provavelmente várias outras coletando dados), portanto, deverá utilizar amostras. Usando os dados da amostra, você ajustará um modelo estatístico para testar suas previsões, ou seja, detectar o efeito de interesse. A estatística resume-se a uma ideia simples: dados observados podem ser previstos a partir de algum tipo de modelo mais um erro associado a tal modelo. O modelo é usado (e normalmente o erro associado a ele) para calcular a estatística teste. Se tal modelo pode explicar grande parte da variação nos dados coletados (a probabilidade de obter essa estatística teste é menor do que 5%), você vai inferir que o efeito de interesse realmente existe na população. Se a probabilidade de obter tal estatística teste é mais do que 5%, você poderá concluir que o efeito é muito pequeno para ser detectado. Em vez de confiar na significância, você pode quantificar o efeito na amostra de uma forma padrão como um *tamanho de efeito* e isso poderá ser útil para determinar a importância de tal efeito. Agora, para o SPSS!

1.11 TERMOS-CHAVE QUE DESCOBRIMOS

- Amostra
- Assimetria
- Curtose
- Desvio padrão
- Desvios
- Distribuição amostral
- Distribuição de probabilidade
- Distribuição normal
- Erro do Tipo I
- Erro do Tipo II
- Erro padrão
- Escore-*z*
- Estatística teste
- Hipótese nula
- Hipóteses experimentais
- Distribuição de freqüências
- Histograma
- Intervalo de confiança
- Leptocúrtica
- Média
- Moda
- Modelo linear
- Nível α
- Nível β
- Platicúrtica
- Poder estatístico
- População
- Soma dos erros ao quadrado (SS)
- Tamanho de efeito
- Teste bilateral
- Teste unilateral
- Variação não-sistemática
- Variação sistemática
- Variância

1.12 QUESTÕES DE REVISÃO DO ALEX ESPERTO

Alex Esperto sabe tudo sobre estatística e SPSS. Ele adora fazer perguntas sobre estatística para os outros a fim de mostrar o quanto ele sabe. Que tal irritá-lo e responder corretamente todas as questões?

1. Por que utilizamos amostras? ①
2. O que é a média e como podemos saber se ela é representativa dos nossos dados? ①
3. Qual é a diferença entre desvio padrão e erro padrão? ①
4. O que é uma estatística teste e o que ela nos informa? ①

Material com direitos autorais

5. O que são os erros do Tipo I e do Tipo II ①
6. O que é o tamanho de efeito e como ele é mensurado? ②
7. O que é poder estatístico? ②

Algumas respostas breves podem ser encontradas no arquivo **Answers (Chapter 1).pdf** disponível no *site* www.artmed.com.br.

1.13 LEITURAS COMPLEMENTARES

COHEN, J. Things I have learned (so far). *American Psychologist*. v. 45, n. 12, 1990. p. 1304-1312.

COHEN, J. The earth is round ($p < .05$). *American Psychologist*. v. 49, 1994. p. 997-1003. Alguns artigos maravilhosos do melhor escritor moderno de estatística.

FIELD, A. P., Hole, G. J. *How to design and report experiments*. Londres: Sage, 2003. Sou suspeito para falar, mas acredito que esse livro traça um bom panorama da teoria estatística básica.

WRIGHT, D. B. *First steps in statistics*. London: Sage. 2002. Capítulos 1, 4 e 5 são introduções bem claras sobre amostragem, intervalos de confiança e outras ideias estatísticas importantes.

Material com direitos autorais

2

O AMBIENTE DO SPSS

2.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ①

Espero que o Capítulo 1 não o tenha feito desistir do livro. Se não, o Capítulo 2 lida com outras informações que você precisará saber: o ambiente do SPSS. Existem vários textos excelentes que fornecem uma introdução ao ambiente de uma forma geral dentro do qual o SPSS opera. Os melhores incluem Kinnear e Gray (2000) e Foster (2001). Esses textos valem a pena ser lidos se você não possui familiaridade com computadores e o SPSS. No entanto, eu tenho consideração pela pouca grana de muitos estudantes e assim para tornar este texto útil para aqueles que não têm experiência com o SPSS, esta seção fornece um guia para o ambiente do *software* – mas se você precisa ir mais a fundo então recomendo que veja os textos citados anteriormente e os manuais do SPSS (você pode também achar os arquivos de ajuda (*help*) úteis). Este capítulo trata de janela base do SPSS (o editor de dados – *data editor* e visualizador – *viewer*) e também apresenta como criar variáveis, entrar com dados e ajustar propriedades das variáveis. Terminaremos vendo como carregar arquivos e salvá-los!

2.2 VERSÕES DO SPSS ①

Este livro é baseado principalmente na versão 13.0 do SPSS (pelo menos em termos de diagramas); no entanto, não se deixe impressionar muito pelos números das versões porque a SPSS tem o hábito de lançar "novas" versões regularmente. Embora isso os faça ganhar muito dinheiro e criar um bom mercado para pessoas que escrevem livros sobre o recurso, existem na realidade poucas diferenças entre essas novas versões e a maioria de nós nem irá percebê-las. Ocasionalmente eles apresentam uma grande mudança (a versão 7.0 mostrou uma mudança dramática em relação à versão 6.0 e a versão 10.0 alterou a forma de entrada das variáveis), mas a maior parte do tempo você pode se virar bem com um livro que não cobre explicitamente a versão que você está utilizando (a primeira edição desse livro foi baseado na versão 9, mas pode ser utilizado sem problemas com as versões 7, 8, 10, 11 e 12). Assim esse livro revisado, embora lide com a versão 13.0, poderá ser utilizado com versões anteriores (7.0, 7.5 e 8.0 – existe algumas poucas diferenças entre as versões 7.0, 8.0 e 9.0, mas qualquer diferença óbvia está destacada quando for relevante). Eu também suspeito que ele possa ser utilizado com

Material com direitos autorais

as versões 14, 15 e 16 quando elas aparecerem (embora a SPSS possa decidir mudar tudo só para me contrariar!). Existem várias diferenças em termos da entrada de dados com as versões anteriores a 10 assim se estás utilizando a versão 9 ou anterior em incluí um arquivo denominado de **Field2000(Chapter1).pdf** no *site* www.artmed.com.br. Esse arquivo é uma cópia do Capítulo 1 da primeira edição do livro e irá mostrar como entrar com dados nessas versões mais antigas.

2.3 INICIANDO ①

O SPSS utiliza principalmente duas janelas: o editor de dados – **data editor** (que é o local onde se entra com os dados e se executa as funções estatísticas) e o visualizador – *viewer* (que é onde os resultados de qualquer análise irão ser apresentados).[1] Existem várias janelas adicionais que poderão ser ativadas como o editor de sintaxe – **sintax editor** (veja a Seção 2.6), que permite que se entre com comandos manualmente (em vez de se utilizar os menus baseados em janelas). Para a maioria dos níveis de conhecimento a janela de sintaxe é redundante, pois você pode executar a maioria das análises simplesmente com o mouse. No entanto, existem muitas funções adicionais que podem ser acessadas utilizando-se a sintaxe e loucos que gostam de estatística podem encontrar muitos usos para ela. Apenas para provar que eu sou o mais louco de todos, existem umas poucas seções onde eu irei forçá-lo a usá-la. ☺

Uma vez que o SPSS tenha sido ativado, uma janela inicial aparecerá (veja a Figura 2.1), que permitirá a seleção de várias opções.[2] Se você já tem um arquivo de dados guardado em algum disco que você gostaria de abrir então selecione *Open an existing data source* (Abrir uma fonte de dados existente) clicando no ○ de forma que ele apareça assim: ◉, essa é a opção por omissão (*default*). No espaço embaixo dessa opção existe uma lista dos arquivos abertos recentemente que você poderá selecionar com o mouse. Para abrir o arquivo selecionado clique em OK. Se você quiser abrir um arquivo de dados que não está na lista selecione *More files...* (Mais arquivos) com o mouse e clique em OK. Isso irá abrir a janela de exploração padrão que permitirá a navegação pelos arquivos do computador para tentar encontrar o arquivo desejado (veja a Seção 2.8). Pode, também, ocorrer que você queira abrir algo que não seja um arquivo de dados, por exemplo, um documento *viewer* (visualização) contendo resultados da sua última análise. Você poderá fazer isso selecionando *Open another type of file* (Abrir Outro Tipo de Arquivo) clicando no ○ de forma que ele apareça assim: ◉, e escolhendo um arquivo da lista ou então selecionando *More files…* (Mais arquivos…) e navegando pelo computador. Se você está iniciando uma nova análise (como estamos fazendo aqui) então queremos digitar nossos dados em um novo editor de dados. Para isso precisamos selecionar *Type in data* (Digite os dados) clicando novamente no ○ e então no OK. Isso irá abrir uma janela de edição de dados (*data editor*) em branco.

2.4 O EDITOR DE DADOS ①

A janela principal do SPSS inclui um editor para a entrada dos dados. Essa janela é onde a maioria da ação acontece. No topo dessa janela existe uma barra de menus semelhante a outras que você já deve ter visto em

[1] Em versões do SPSS anteriores a versão 7.0, os gráficos aparecem em uma janela separada conhecida como carrossel gráfico – *chart carousel*); no entanto, as versões 7.0 e posteriores incluem os gráficos na janela de saída, que é denominada de navegador de saída – *output navigator* (versão 7.0) e o visualizador de saída – *output viewer* (versão 8.0 e posteriores).

[2] De fato, essa janela não aparece em versões anteriores a 10.0. Em vez disto, um editor de dados em branco é carregado e você poderá abrir os arquivos utilizando os menus (veja a seção 2.8 ou o arquivo **Field2000(Chapter1).pdf** no *site* www.artmed.com.br).

Material com direitos autorais

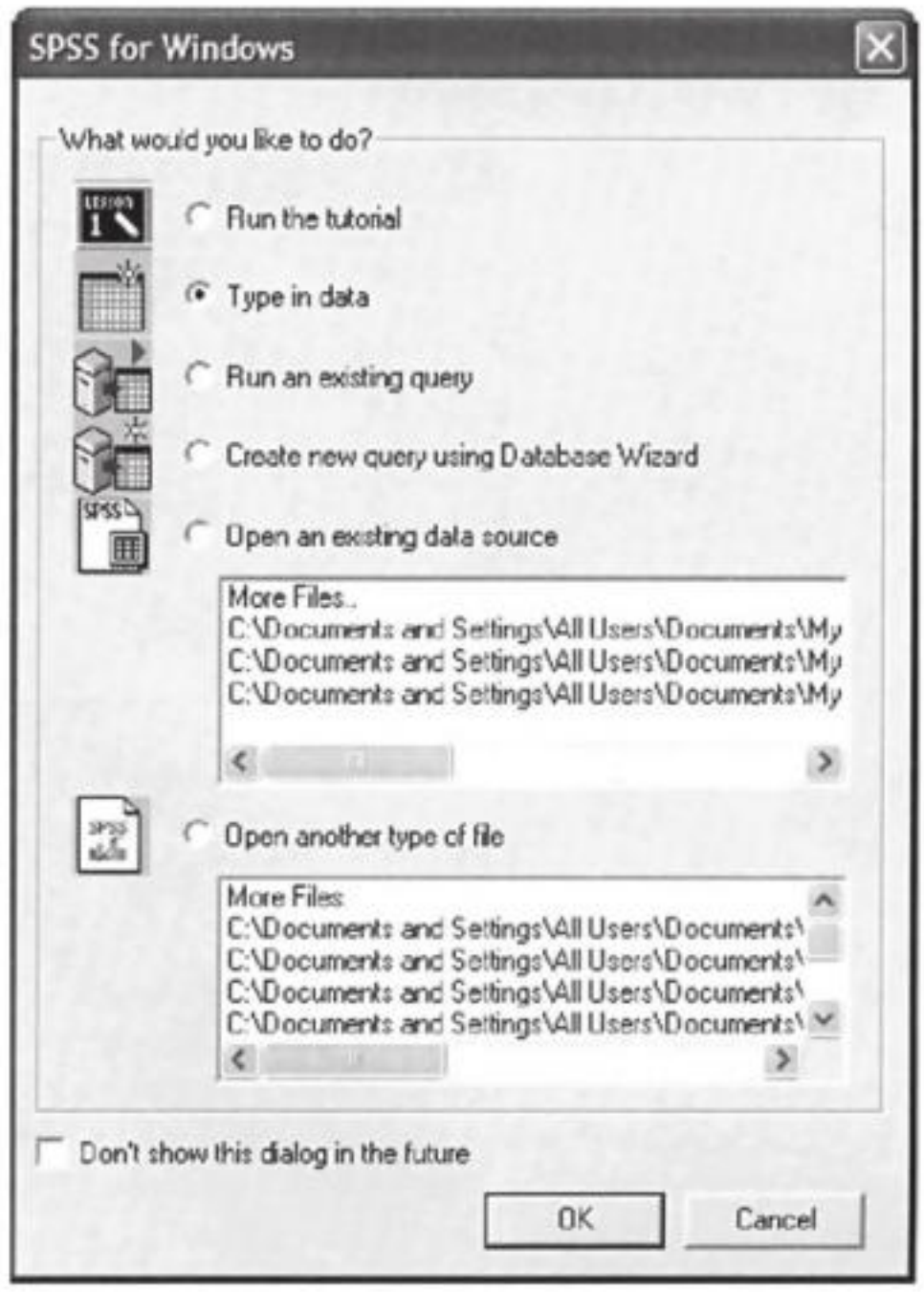


Figura 2.1 A janela inicial do SPSS.

outros programas (tal como o Word). A figura 2.2 mostra essa barra de menu e o editor de dados. Existem vários menus no topo da tela (por exemplo, *File* (Arquivo), *Edit* (Editar), etc.) que podem ser ativados pela utilização do mouse para mover o cursor sobre o menu desejado e então pressionar o botão esquerdo uma vez (pressionar esse botão é normalmente conhecido como 'clicar'). Quando você clica em um menu, uma janela se abre mostrando uma lista de opções que podem ser ativadas movendo o cursor para a opção desejada e então clicando-a. Frequentemente selecionar uma opção do menu faz que uma nova janela apareça, essas janelas são denominadas de "caixas de diálogo" (*dialog boxes*). Quando me referir a opções selecionáveis de um menu eu irei representar a ação utilizando negrito com uma seta indicando o caminho do mouse (assim, cada seta representa colocar o cursor sobre a palavra e clicar com o botão esquerdo do mouse). Por exemplo, se eu quiser dizer que você deve selecionar a opção *Save As...* (Salvar Como) do menu *File* (Arquivo), escreverei isso como **File⇒Save As...** (**Arquivo⇒Salvar Como**).

O editor de dados tem dois painéis (*views*): o de **dados** e o de **variáveis**. O editor de dados é para entrar obviamente dados e o editor de variáveis permite que definamos várias características das variáveis do editor de dados.[3] Na

[3] Antes do SPSS versão 10, a entrada de dados era feita utilizando um único painel (veja Field, 2000, Capítulo 1 que está disponível *site* www.artmed.com.br como **Field2000(Charpter1).pdf).**

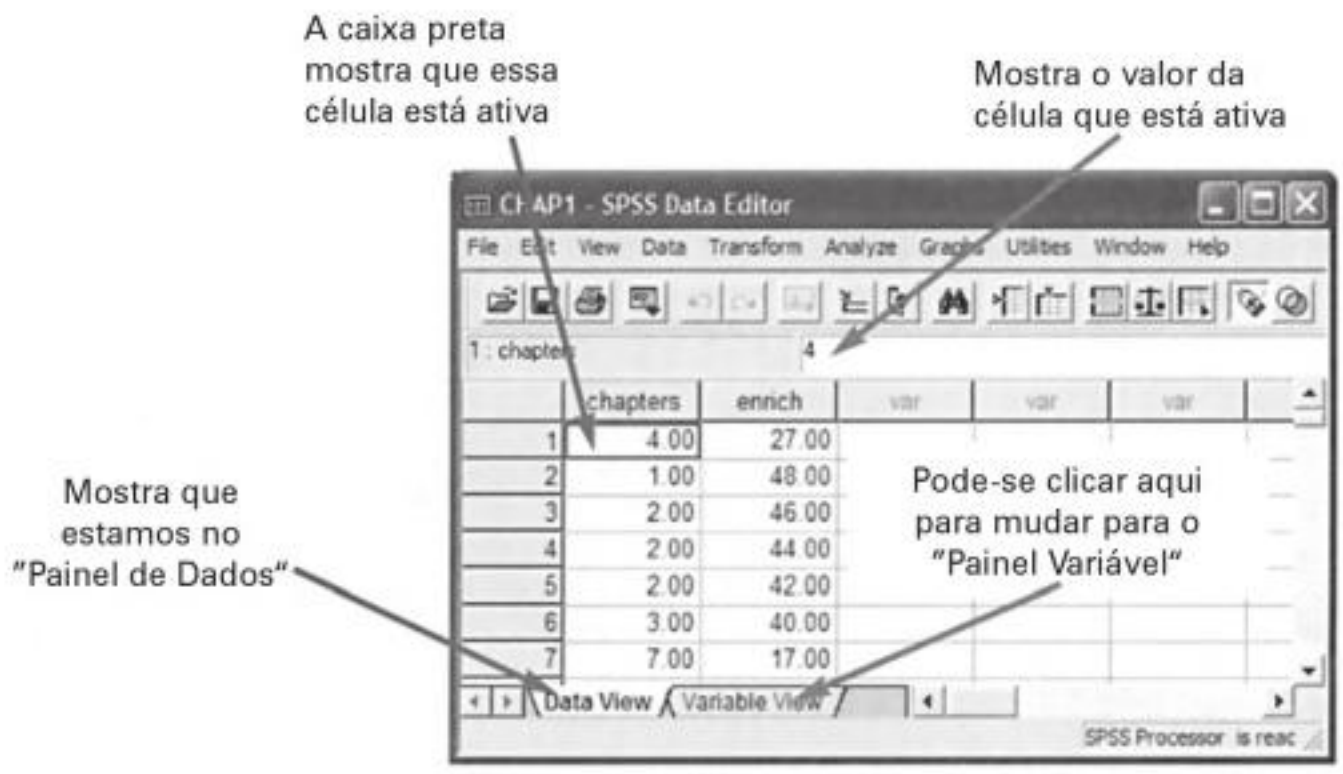


Figura 2.2 O editor de dados do SPSS (versão 10 em diante).

parte de baixo do editor de dados, você pode observar que existem duas orelhas (*tabs*) denominadas como *Data View* (Painel de Dados) e *Variable View* (Painel de Variáveis) (Data View / Variable View) e tudo o que temos que fazer para passar de um para outro é clicar com o mouse. Se estivermos no editor de dados então essa orelha terá um fundo branco e o painel das variáveis será cinza e se estivermos no painel das variáveis então o oposto irá ocorrer. Primeiro, vamos dar uma olhada em algumas características do editor de dados, essas são características que não se alteram quando nós mudamos de um painel para outro. Antes, vamos verificar os menus.

Em vários *softwares* dentro dos menus serão encontradas algumas letras sublinhadas, isso representa um "atalho de teclado" para facilitar o acesso a tal função. É possível selecionar muitas funções sem a utilização do mouse, e um usuário experiente pode achar esses atalhos mais rápidos que posicionar o mouse sobre o local adequado na tela. As letras sublinhadas nos menus indicam que a opção pode ser ativada pressionando simultaneamente *Alt* no teclado seguido da letra sublinhada. Assim para acessar a opção *Save As...* (Salvar Como), utilizando apenas o teclado, você deve pressionar *Alt* e F simultaneamente (que irá ativar o menu *File* – Arquivo)*; então, mantendo a tecla *Alt* pressionada, pressione A que é letra sublinhada em *Save As...*. No SPSS 13.0 (pelo menos quando estiver rodando no Windows XP), essas letras sublinhadas não são visíveis, no entanto, você ainda pode utilizar os atalhos de teclado. Se você pressionar *Alt* então as letras sublinhadas se tornarão visíveis, e uma vez pressionada a tecla *Alt* você pode apenas pressionar as letras sublinhadas para navegar pelos menus. Desta forma, ao longo do livro, eu incluirei as letras sublinhadas para os leitores que quiserem utilizar os atalhos de teclado, mas para aqueles, exceto você, que não querem, tentem não ser desestimulados pelas letras sublinhadas no livro!

Abaixo um guia breve de referência para cada um dos menus e algumas das opções que eles contêm. Isso é meramente um resumo e descobriremos as características de cada menu conforme formos progredindo ao longo do texto.

- ***File* (Arquivo):** Esse menu permite fazer coisas gerais tais como salvar arquivos, grá-

* N. de T. Evidentemente, se o software estiver traduzido, para se ativar o menu *Arquivo*, será necessário pressionar *Alt* + *A*.

ficos ou saídas de análise. Da mesma forma, pode-se abrir arquivos já salvos e imprimir gráficos, dados ou saídas. Em resumo, ele contém todos as opções costumeiramente encontrados no menu *File* (Arquivo).

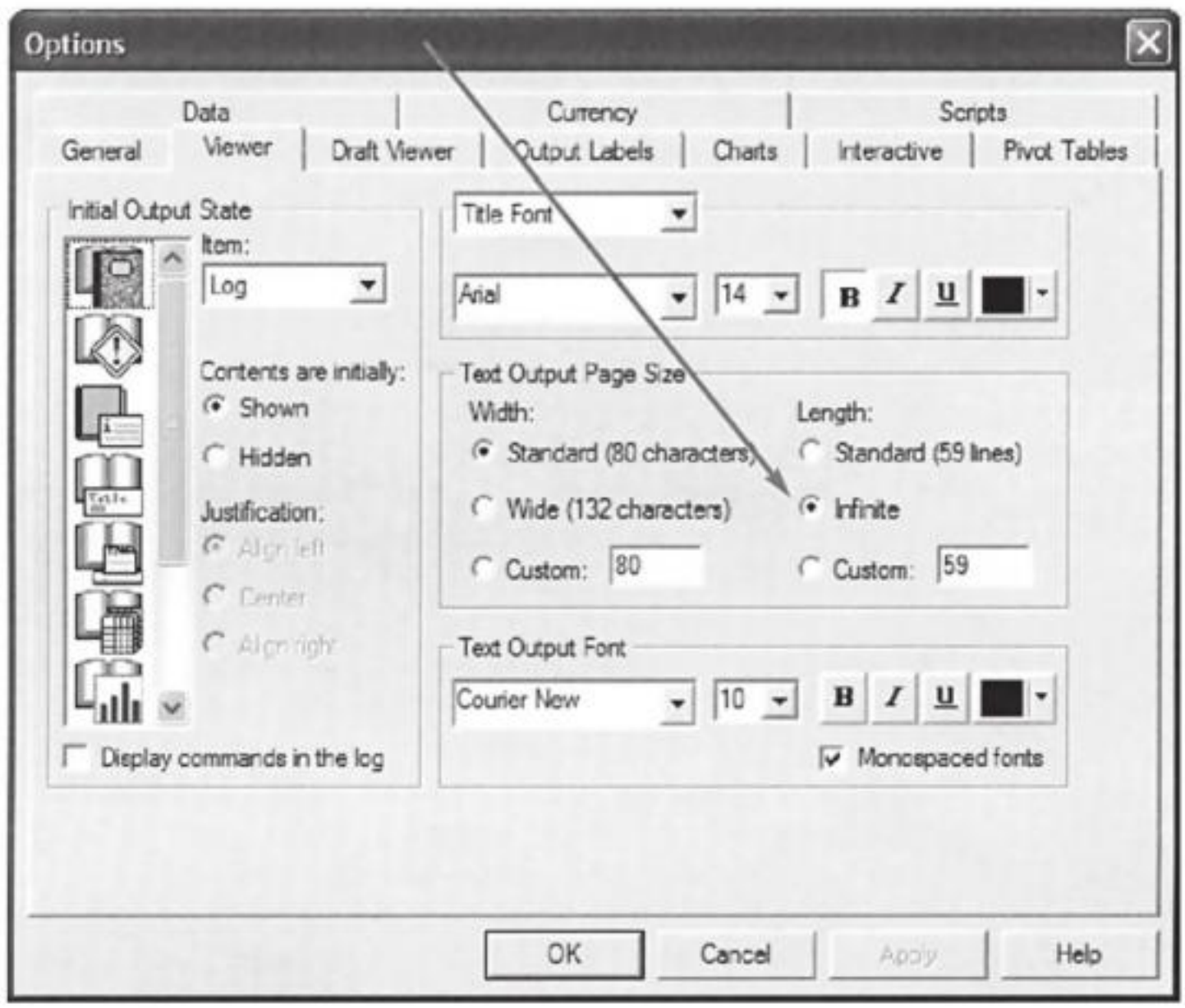
- **Edit (Editar):** Esse menu contém funções de edição para o editor de dados. No SPSS para Windows é possível recortar e colar blocos de números de uma parte do editor para outro (que pode ser bastante útil quando você percebe que digitou montes de números no lugar errado). Você pode, ainda, utilizar *Options* (Opções) para selecionar várias preferências tais como as fontes que são utilizadas nas saídas. As preferências por omissão (*default*) são boas para a maioria das situações; a única coisa que você pode querer mudar é o tamanho da página de saída do visualizador colocando a opção como infinito (isso salvará muitas árvores quando você começar a imprimir coisas). Para fazer isso você precisa selecionar a orelha *Viewer* (Visualizador) no início da janela.
- **View (Exibir):** Esse menu lida com especificações do sistema tais como se o editor apresentará linhas de grade, ou se você apresentará os rótulos (*labels*) dos valores (exatamente o que são rótulos se tornará claro mais tarde).
- **Data (Dados):** Esse menu permite que se façam alterações no editor de dados. As características importantes são *insert variable* (inserir variável), que é utilizada para inserir uma nova variável no editor de dados (isto é, adicionar uma nova coluna); *insert case* (inserir caso), que é utilizado para inserir uma nova linha de dados entre duas já existentes, *split file* (dividir arquivo), que é utilizado para dividir um arquivo por uma variável de agrupamento (veja a Seção 3.4) e *select cases* (selecionar casos), que é utilizada para rodar uma análise utilizando somente uma amostra dos casos (valores).
- **Tranform (Transformar):** Você deve utilizar esse menu se quiser manipular uma das variáveis de alguma forma. Por exemplo, você pode utilizar *recode* (recodificar) para mudar os valores de certas variáveis (por exemplo, se você quiser atribuir nomes diferentes por alguma razão) – veja o Quadro 3.2. A função *compute* (calcular) é também útil para transformar dados (por exemplo, para criar uma nova variável que é a média de duas existentes). Essa função permite que se execute qualquer número de modificações nas variáveis existentes (veja a Seção 3.3.4).
- **Analyze (Analisar):** A diversão começa aqui, porque todos os procedimentos estatísticos estão nesse menu.[4] A seguir, há um breve guia das opções disponíveis que serão utilizadas durante o curso deste livro (isso é apenas uma pequena porção do que está disponível):
  - **(a) Descriptive Statistics (Estatística Descritiva):**[5] Esse menu serve para determinar medidas descritivas (média, medidana, moda, etc.), frequências e para uma exploração geral dos dados. Existe ainda um comando denominado de *crosstabs* (tabulação cruzada) que é útil para explorar frequências de dados e executar testes tais como o qui-quadrado, o exato de Fisher e o *kappa* de Cohen.
  - **(b) Compare Means: (Comparar Médias):** Aqui é onde se podem encontrar os testes *t* (para amostras relacionadas e independentes – Capítulo 7) e a ANOVA independente (Capítulo 8).
  - **(c) General Linear Model: (Modelo Linear Generalizado):**[6] Esse menu é para realizar Análises de Variâncias complexas tais como a de dois fatores (relacionada, não-relacionada ou mista), a de um fator com medidas repetidas e a análise de variância multiva-

[4] Esse menu é denominado de **Statistics** (Estatística) na versão 8.0 ou anterior.

[5] Esse menu é denominado de **Sumarize** (Resumir) na versão 8.0 ou anterior.

[6] Esse menu é denominado de **ANOVA Models** (Modelo ANOVA) na versão 6 do SPSS.

Material com direitos autorais



riada (MANOVA) – veja os capítulos 9, 10, 11, 12 e 14.

(d) Correlate: (Correlação): Não precisa ser um gênio para perceber que é aqui que as técnicas de correlação são encontradas! Você poderá executar correlações bivariadas tais como o cálculo do r de Pearson, o rô de Spearman (ρ) o tau de Kendall (τ) bem como as correlações parciais (veja o capítulo 4).

(e) Regression: (Regressão): Existe uma variedade de técnicas de regressão disponíveis no SPSS. Você poderá executar uma regressão linear simples, uma múltipla (Capítulo 5) e técnicas mais avançadas tais como a regressão logística (Capítulo 6).

(f) Loglinear: (Loglinear): A análise loglinear está escondida nesse menu, esperando por você, e pronta para atacar como uma tarântula de sua toca (Capítulo 16).

(g) Data Reduction: (Redução de dados): Você encontrará a Análise de Fatores aqui (Capítulo 15).

(h) Scale: (Escala): Aqui será encontrada a Análise de Confiabilidade (Capítulo 15).

(i) Nonparametric Tests: (Testes não-paramétricos): Existem uma variedade de estatísticas não-paramétricas disponíveis tais como o a estatística de aderência pelo qui-quadrado, o teste binomial, o teste de Mann-Whitney, o de Kruskal-Wallis, Wilcoxon e a ANOVA de Friedman (Capítulo 13).

- **Graphs (Gráficos):** O SPSS vem com seu próprio módulo gráfico bastante versátil. Entre os tipos de gráficos disponíveis podem ser incluídos: diagramas de barras, histogramas, diagramas de dispersão, dia-

gramas de caixa-e-bigodes, diagramas de *pizza* e diagramas de barras de erros para enumerar apenas alguns. Existe também a facilidade para editar qualquer tipo de gráfico para fazê-lo parecer bastante atrativo, que, na minha opinião, é bem formulado.

- **Window (Janela):** Permite que se mude de uma janela para outra. Assim, se você estiver olhando para uma saída e quiser retornar para a planilha, você poderá fazê-lo utilizando esse menu. Existem ícones como atalhos para a maioria das opções desse menu assim ele não é particularmente útil.
- **Help (Ajuda):** Esse é um menu inestimável pois ele oferece a você ajuda on-line tanto sobre o próprio sistema quanto sobre testes estatísticos. Embora os arquivos de auxílio estatístico possam ser, às vezes, praticamente inúteis (afinal de contas o programa não foi projetado para ensinar estatística), eles certamente não são um substituto para a aquisição de um bom conhecimento, mas poderá livrar você do aperto em algumas situações.

Assim como os menus, existe ainda um conjunto de ícones no topo da janela do editor de dados (veja a Figura 2.2) que são atalhos para recursos específicos que são utilizados com frequência. Todos esses recursos podem ser acessados pelo sistema de menus, mas utilizando os ícones você poupará tempo. Abaixo uma lista breve desses ícones e para que servem:

Esse ícone fornece a opção de abrir arquivos salvos previamente (se você estiver no editor de dados o SPSS assume que você quer abrir um arquivo de dados; se você estiver no visualizador de saídas, ele assumirá que você quer abrir um arquivo desse tipo).

Esse ícone permite que um arquivo seja salvo. Ele irá salvar o arquivo corrente, isto é, em que você estiver trabalhando quando clicar no ícone (seja ele de dados ou de saída). Se o arquivo ainda não tiver sido previamente salvo ele abrirá a caixa de diálogos *save data as* (salvar dados como).

Aqui será aberta a caixa de diálogos de impressão não importa o tipo de arquivo que você estiver trabalhando (dados ou saídas). As opções de impressão irão depender da impressora que você estiver utilizando. Uma dica útil é selecionar partes da saída clicando em porções da janela de saída (veja a Seção 2.5). Quando a caixa de diálogos de impressão aparecer lembre-se de clicar na opção de impressão apenas do texto selecionado. Selecionando partes do texto irá salvar muitas árvores pois a opção do SPSS é imprimir tudo o que estiver na janela de saída.

Clicando nesse ícone ativará a lista das últimas 12 caixas de diálogo que foram utilizadas. A partir dessa lista você poderá selecionar qualquer caixa da lista que ela aparecerá na tela. Esse ícone facilita a repetição de partes de uma análise.

Esse ícone permite que se vá direto a um caso (isto é, um participante). Ele será útil se estivermos trabalhando com grandes arquivos de dados. Por exemplo, se você quiser analisar um levantamento com 3000 respondentes será bastante tedioso ir rolando a tela até encontrar a resposta de um participante em particular. Esse ícone pode ser utilizado para pular diretamente para o caso (por exemplo, caso 2407). Clicando nesse ícone ativará uma caixa de diálogos que irá requerer que seja digitado o número do caso procurado.

Clicar nesse ícone fornecerá informações sobre uma determinada variável do editor de dados (uma caixa de diálogos permite a escolha de qual variável se poderá solicitar um resumo das informações).

Esse ícone não é para permitir que você faça observação de pássaros, mas ele permite que se procure por palavras ou números no arquivo de dados ou na janela de saída.

Ao clicar nesse ícone um novo caso será inserido no editor de dados (assim, ele

Material com direitos autorais

cria uma linha em branco no ponto em que ele estiver destacado – cursor estiver posicionado). Essa função é muito útil se for necessário adicionar novos dados ou se você esqueceu-se de colocar os dados de um participante em particular no editor de dados.

Clicar aqui irá criar uma nova variável à esquerda da variável que está ativa no momento (para ativar uma variável, simplesmente clique uma vez no nome dela no topo da coluna).

Esse ícone é um atalho para o comando **Data⇒Split File...** (Dados⇒Dividir Arquivo) (veja a Seção 3.4). Cientistas sociais às vezes conduzem experimentos com diferentes grupos de pessoas. No SPSS diferenciamos grupos de pessoas pela utilização de variáveis de código (veja a Seção 2.4.4) e essa função permite que se divida nossa saída de acordo com essa variável. Por exemplo, podemos querer testar a habilidade estatística de homens e mulheres. Podemos codificar cada participante com um número que representa o gênero (por exemplo, 1 = mulher, 0 = homem). Se quisermos então saber a habilidade estatística média de cada sexo nós simplesmente dividimos o arquivo utilizando à variável **gênero**. Qualquer análise subseqüente será executada para homens e mulheres separadamente.

Esse ícone é um atalho para a função **Data⇒Weight Cases...** (Dados⇒Ponderar Casos). Essa função é necessária para quando trabalharmos com dados na forma de freqüências (veja a Seção 16.4.2) e é útil para algumas situações avançadas de amostragem.

Esse ícone é um atalho para a função **Data⇒Select Cases...** (Dados⇒Selecionar Casos). Se você quer analisar apenas uma parte dos dados, essa é a opção a ser utilizada. Essa função permite que se especifique que casos devem ser incluídos na análise.

Ao clicar nesse ícone os rótulos dos dados para qualquer variável que tenha esse tipo de característica aparecerão ou serão ocultados. Algumas vezes agrupamos casos e para isso utilizamos uma variável de grupo ou codificadora para informar o computar a que grupo cada caso pertence. Por exemplo, se codificamos o gênero como 1 = mulher e 0 = homem então o computador saberá que cada vez que o valor 1 aparecer na coluna (variável) **gênero** a pessoa é uma mulher. Se você pressionar esse ícone, a legenda (mulher) irá aparecer no editor de dados ao invés do valor numérico (1); assim, você irá ver as palavras *homem* e *mulher* na coluna **gênero** ao invés de uma série de números (0 ou 1). A ideia se tornará mais clara na Seção 2.4.4.

2.4.1 Entrando com os dados no editor ①

Quando você executa o SPPS pela primeira vez ele vai apresentar uma planilha em branco no editor de dados e essa planilha será nomeada de "sem título" (isso é ridículo pois uma vez que ela foi nomeada de "sem título" ela já terá um título!). Quando entrar com um novo conjunto de dados, você deve entrar com os seus dados de uma forma lógica. O editor do SPSS é organizado de forma que *cada linha representa os dados de um objeto (indivíduo) enquanto que cada coluna representa uma variável*. Não existe uma separação para variáveis dependentes e independentes: os dois tipos devem ser colocados em colunas separadas. O ponto chave é que cada linha representa os dados de uma pessoa. Portanto, qualquer informação sobre cada caso deve ser fornecido com o editor de dados. Por exemplo, imagine que estamos interessados em investigar as diferenças entre os sexos quanto a percepção de dor criada por estímulos de quente e frio. Você pode colocar as mãos de algumas pessoas em um recipiente com água gelada por um minuto e perguntar a elas para atribuirem um valor do quanto doloroso isso foi em uma escala de 0 a 10. Você então pode pedir a elas

Material com direitos autorais

para segurar uma batata quente e novamente medir as percepções de dor. Imagine que eu sou um participante. Você terá uma única linha representando os meus dados, assim existirá uma coluna diferente para o meu nome, minha idade, meu gênero, minha percepção de dor com a água gelada e minha percepção de dor com a batata quente: Andy, 31 (quando esse livro apareceu), homem, 7, 10.

A coluna com a informação sobre o meu gênero é uma variável de grupo: posso pertencer ou ao grupo dos homens ao das mulheres, mas não a ambos. Como tal, essa variável é uma variável entre grupos (pessoas diferentes pertencem a grupos diferentes). Em vez de representar pessoas com palavras, no SPSS utilizamos números. Isso envolve atribuir a cada grupo um número e então informar ao SPSS que número representa cada grupo. No entanto, variáveis entre grupos são representadas por uma única coluna na qual o grupo que a pessoa pertence é definido por um número (veja a Seção 2.4.4). Por exemplo, podemos decidir que uma pessoa é homem e daremos a ela o número 0, ou se ela for mulher será atribuído o número 1. Nós teremos então que informar ao SPSS que cada vez que ele encontrar um número 1 em uma dada coluna a pessoa é mulher e se o valor for 0 essa pessoa é homem. Variáveis que especificam a que grupo entre vários uma pessoa pertence pode ser utilizada para dividir arquivos (assim, no exemplo da dor você pode executar uma análise com os homens e as mulheres separadamente – veja a Seção 3.4).

Finalmente, as duas medidas da dor são medidas repetidas (todos os participantes foram sujeitos aos estímulos de quente e frio). Desta forma, os níveis dessa variável podem ser fornecidos em colunas separadas (uma para a dor ao estímulo quente e outra para o frio).

O editor de dados é formado de células, que são apenas quadros em que os valores dos dados podem ser colocados. Quando uma de tais células está ativa ela se torna mais clara com as margens em preto destacadas (veja a Figura 2.2). Você pode se movimentar pelo editor de dados, de uma célula para outra, utilizando as teclas de setas ← ↑ ↓ → (encontradas à direita no teclado) ou clicando com o mouse na célula que deseja ativar. Para entrar com um valor no editor simplesmente mova-se para a célula que deseja colocar o valor, digite o valor, e então pressione a tecla com a seta para o lado que deseja continuar indo. Assim, para entrar com uma linha de dados, mova-se para a esquerda na mesma linha, digite o valor e então pressione → (esse processo entra com o valor e então movimenta o cursor para a próxima célula à direita).

2.4.2 Criando uma variável ①

O primeiro passo para entrar com dados é criar alguma variável utilizando o painel "Variáveis" do editor de dados e então entrar com os dados no painel "Dados" do editor. Nós seguiremos por esses dois passos trabalhando com um exemplo. Imagine que estamos interessados em verificar a diferença entre professores e alunos. Tomamos ao acaso uma amos-

Dica da Samanta Ferrinho

Para resumir, qualquer variável medida sobre os mesmos participantes (uma medida repetida) deve ser representada por várias colunas (cada coluna representando um nível da variável de medidas repetidas). Entretanto, qualquer variável que define grupos diferentes de pessoas (como ocorre quando o delineamento entre grupos é utilizado e diferentes participantes são atribuídos a diferentes níveis da variável independente) é definida utilizando uma única coluna. Essa ideia se tornará mais clara a medida que você aprender como executar procedimentos específicos.

Material com direitos autorais

tra de cinco professores e alunos de psicologia da universidade de Sussex e verificamos quantos amigos cada um tem, o consumo semanal de álcool (em alguma unidade), a renda anual e quão neuróticos eles são (escores mais altos significa maior neurose). Esses dados estão na Tabela 2.1.

2.4.3 O visualizador de variáveis (*Variable View*) ①

Antes de realmente entrarmos com os dados no editor nós precisamos criar as variáveis. Para criar as variáveis utilizamos o painel variáveis do editor de dados. Para acessar esse painel clique na orelha *Variable View* no fundo do editor de dados (Data View / Variable View); a janela irá mudar e a orelha do visualizador ficará branco (Figura 2.3).

Cada linha do editor de variáveis representa uma variável, e para particularizar uma variável basta configurar a coluna adequada. Você pode alterar várias características tais como, a largura da coluna, o número de casas decimais após a vírgula que será mostrado e se os valores ficarão alinhados à direita, à esquerda ou no centro (mexa um pouco com esses valores e você pegará o jeito). As características mais importantes são as seguintes:

Name Você pode colocar nessa coluna o nome para cada variável. Esse nome irá aparecer no início da coluna quando você estiver no painel de dados e serve para ajudar você a identificar as variáveis no painel dos dados. Existem algumas regras gerais sobre nomes de variáveis tais como o de eles só poderem ter oito caracteres ou menos, as letras só podem ser as minúsculas e você não poderá utilizar espaços em branco. Se for violada qualquer uma dessas regras o computador dirá que o nome da variável é inválido quando você clicar em uma célula diferente, ou tentar se movimentar para fora da célula utilizando uma tecla de seta.

Type Você pode ter diferentes tipos de dados. O tipo mais comum é *numeric* (numérico) (esse é o padrão, isto é, atribuído por omissão e apenas informa que essa é uma variável que contem números). Outro tipo é a variável *string* (texto), que é uma contração do "coleção de letras". Assim se você quiser digitar o nome das pessoas, por exemplo, você precisará alterar o tipo de variável para *string* (texto) em vez de *numeric* (numérica).

Label O nome de uma variável (veja acima) é restrito a oito caracteres, assim o SPSS fornece uma maneira para que possamos identificar as variáveis de uma forma mais conveniente (esse rótulo poderá ter maiúsculas e mesmo espaços!). Isso pode não parecer importante mas é provavelmente um dos melhores hábitos que você poderá adquirir. Se você tiver uma variável chamada "número

Tabela 2.1 Alguns dados para brincar

	Amigos	Álcool	Renda	Neurose
Professor	5	10	20000	10
Professor	2	15	40000	17
Professor	0	20	35000	14
Professor	4	5	22000	13
Professor	1	30	50000	21
Aluno	10	25	5000	7
Aluno	12	20	100	13
Aluno	15	16	3000	9
Aluno	12	17	10000	14
Aluno	17	18	10	13

Material com direitos autorais

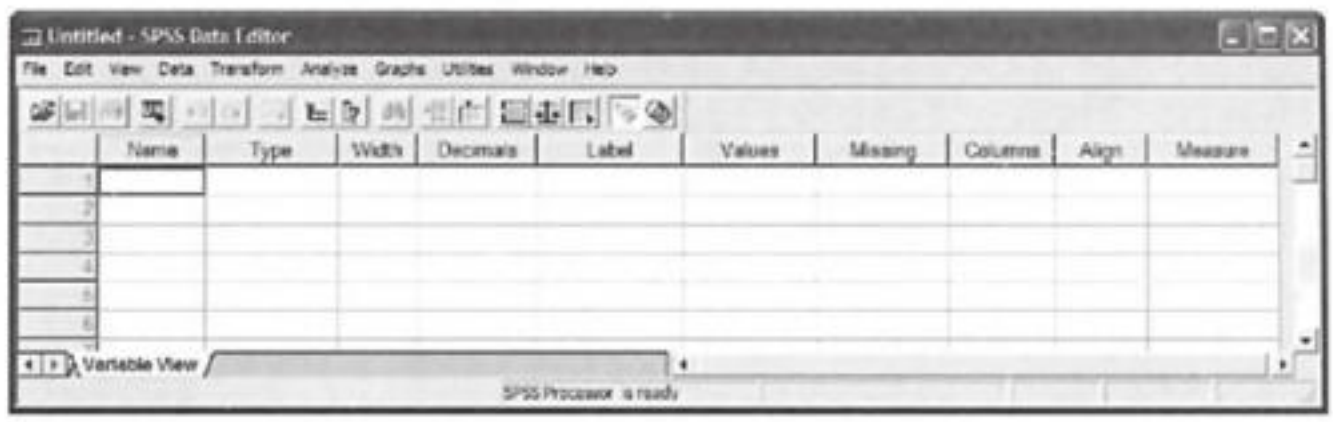


Figura 2.3 O painel "visualizador de variáveis" do editor de dados do SPSS.

de vezes que eu quis me dar um tiro durante uma aula de estatística do Andy Field", você poderá denominá-la de "tiro". Se você não adicionar um rótulo, o SPSS irá utilizar o nome da variável (nesse caso tiro) em todas as saídas de uma análise. Isso está bem e poderá ser bom, mas acredite, algum tempo depois quando você tentar rever os dados e as análises provavelmente irá pensar "que raios significa tiro?" ou "que diabos é essa variável sftg45c?". Assim, mantenha o bom hábito de rotular as suas variáveis!

Values Essa coluna é para atribuir números para representar grupos de pessoas (veja a Seção 2.4.4).

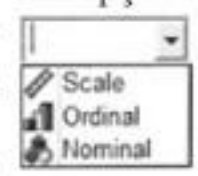
Missing Essa coluna e para atribuir valores a dados *missing* (que faltam ou não foram obtidos) (veja a Seção 2.4.6).

Measure Aqui é onde se define o nível na qual a variável foi mensurada (*Nominal, Ordinal* ou *Scale*) (Nominal, Ordinal ou Numérica).

Para entrar com os dados anteriores no editor do SPSS você precisará criar várias variáveis. Se começarmos com a variável **número de amigos**, devemos seguir esses passos:

1. Clicar com o mouse na primeira célula em branco da coluna denominada *Name* (Nome).
2. Digite a palavra *friends* (amigos) lembre-se que só podemos utilizar nomes com oito ou menos caracteres assim não será possível digitar *number of friends* (número de amigos).
3. Saia da célula utilizando as teclas com as setas do teclado (ou você pode apenas clicar em uma célula diferente, mas com as setas é mais rápido).

Você acaba de criar sua primeira variável! Você deve ter notado que uma vez que digitou um nome o SPSS cria alguns valores padronizados (por omissão) para a variável (tal como assumir que ela é numérica e colocar duas casas após a vírgula). Agora em virtude de que você quer manter bons hábitos, mova para a célula na coluna chamada *Label* (Rótulo) e digite 'número de amigos'. Finalmente, podemos especificar o nível em que a variável foi mensurada (veja o Quadro 2.1) indo para a coluna rotulada *Measure* (Medida) e selecionando uma das opções *Nominal,Ordinal* ou *Scale* (Nominal, Ordinal ou Numérica) da lista suspensa (veja o Quadro 2.1)



Uma vez que uma variável foi criada, você pode retornar para o painel dos dados clicando na orelha *Data View* (visualizador de dados) na parte de baixo do editor de dados (Data View Variable View). O conteúdo da janela irá mudar e você poderá notar que a primeira coluna apresenta agora o rótulo *friends* (amigos). Para entrar com os dados, clique na célula no topo da coluna denominada *friends* (amigos) e digite o primeiro valor, 5. Para registrar esse valor na célula, você precisa se movimentar para uma célula diferente e em virtude de estarmos inserindo dados em colunas, a maneira mais fácil de fazer isso é pressionando a tecla ↓. Essa ação moverá você uma posição abaixo e o número 5.00 deverá

Quadro 2.1

Níveis de Medida ①

Existem vários níveis nas quais as variáveis podem ser medidas: nominal, ordinal, intervalar e de razão. O mais baixo é o **nominal**, onde os números apenas representam nomes. Por exemplo, se perguntarmos às pessoas se ler este capítulo as deixa com tédio, as respostas serão *sim* ou *não*. Assim, as pessoas serão classificadas em duas categorias: entediadas ou não-entediadas, obviamente aqui o valor do número é irrelevante e ele meramente representará uma das duas categorias (se você estiver utilizando uma variável codificadora [próxima seção] então os dados são *nominais*). Uma variável **ordinal** nos fornece mais informações do que uma nominal. Se utilizarmos uma escala ordinal para medir algo, podemos informar não apenas que coisas ocorreram, mas também a ordem em que elas aconteceram. No entanto, esses dados não nos informam o valor das diferenças entre as categorias. Por exemplo, se classificarmos três pessoas de acordo com o quanto elas estão entediadas – muito entediada, tédio médio e pouco entediada. Esses rótulos nos informam sobre o nível de tédio. Ao utilizar categoriais ordinais nós agora sabemos que a pessoa mais entediada apresenta mais tédio do que a pessoa menos entediada! Dados em **intervalo** são escores medidos por uma escala em que todos os intervalos são iguais. Por exemplo, em vez de perguntar às pessoas se elas estão entediadas, poderíamos medir o tédio com uma escala de 10 pontos (com 0 representando alguém muito interessado e 10 alguém totalmente entediado). Para dados em intervalo um aumento no tédio representado por uma mudança do grau 3 na escala para 4 tem o mesmo valor que um aumento do grau 9 para o 10, mas não é verdade que um entediado de grau 9 apresenta um tédio 3 vezes maior do que um classificado com grau 3. Dados de **razão** apresentam essa propriedade mas aqui é possível afirmar que alguém com um grau 8 de tédio está duas vezes mais entediado do que alguém que foi classificado com grau 4 de tédio. Esses dois tipos de dados são representados pela opção *Scale* (numérica) no SPSS. Deve parecer óbvio que em algumas ciências sociais, principalmente psicologia, é extremamente difícil estabelecer se um dado é, de fato, intervalar (nós podemos realmente dizer que uma alteração na escala de tédio representa uma mudança genuína na experiência de apresentar tédio?) – veja Field e Hole (2003).

aparecer na célula acima. Entre com o próximo número, 2, e então pressione ↓ novamente para seguir adiante.

2.4.4 Criando variáveis codificadas ①

Nas seções anteriores eu mencionei as variáveis de grupo (codificadoras) e esta seção é dedicada uma descrição completa desse tipo de variável (é um tipo de variável que será muito usada). Uma variável codificadora (também conhecida como variável de grupo ou agrupadora) é uma variável que consiste em uma série de números representada em níveis de uma variável tratamento ou que descreve diferentes grupos de pessoas. Em experimentos, variáveis de grupo são utilizadas para representar variáveis independentes que tenham sido mensuradas entre grupos (isto é, diferentes participantes foram atribuídos a grupos diferentes). Assim, se você está realizando um experimento com um grupo de participantes em uma condição experimental e um grupo diferente de participantes em um grupo em um grupo-controle, você poderá atribuir ao grupo-experimental o código 1 e ao grupo controle o código 0, por exemplo. Quando você entrar com os dados no editor, então deverá criar uma variável (que poderá chamá-la de **grupo**) e digitar 1 para todos os participantes do grupo-experimental e 0 para qualquer participante do grupo-controle. Esses códigos informam ao computador que todos os casos que foram atribuídos ao grupo 1 devem ser tratados como se pertencentes a um mesmo grupo e da mesma forma para os casos em que foram

Material com direitos autorais

Existe uma regra simples de como as variáveis devem ser colocadas no editor de dados do SPSS: níveis de variáveis entre grupos devem variar verticalmente (ocupar uma única coluna), enquanto que variáveis dentre-assuntos (medidas repetidas) devem variar horizontalmente (mesma linha). Veremos como essa regra funciona no Capítulo 6.

atribuídos o valor 0. Em situações que não sejam experimentos, você poderá simplesmente utilizar os códigos para distinguir grupos de pessoas ocorrendo naturalmente (por exemplo, você pode atribuir aos alunos o código 1 e aos professores o código 0).

Nós temos uma variável de código em nossos dados: é aquela descrevendo se uma pessoa é um professor ou um aluno. Para criar essa variável codificadora, seguimos os passos que seriam dados para criar uma variável normal, mas devemos também informar ao computador que códigos numéricos foram atribuídos e para quais grupos. Assim, antes de mais nada, retorne ao painel de visualização das variáveis se você ainda não estiver nele e mova-se para a célula na segunda linha do editor na coluna rotulada como *Name* (Nome). Digite um nome (chame-a de **grupo**). Eu ainda estou tentando fazer com que você adquira bons hábitos, assim mova-se ao longo da segunda linha para a coluna denominada de *Label* (Rótulo) e forneça a variável uma descrição pormenorizada tal como 'A pessoa é professor ou aluno?' Então para definir os códigos dos grupos, mova-se ao longo da segunda linha para a coluna denominada de Values. Você verá que na célula está escrito *None* (Nada) e que existe um botão (veja a Figura 2.4). Clique nesse botão para acessar a caixa de diálogos *Value labels* (Valores dos rótulos) (veja a Figura 2.4).

A caixa de diálogos *Value labels* (valores dos rótulos) é utilizada para especificar o código dos grupos. Isso pode ser feito em passos simples. Primeiro clique com o mouse no espaço em branco próximo de onde diz *Value* (Valor) (ou pressione *Alt* e *u* ao mesmo tempo) e digite o código (por exemplo, 1). Esses códigos podem ser completamente arbitrários; por simples convenção as pessoas normalmente utilizam os valores 0, 1, 2, 3, etc., mas na prática você pode colocar um valor de 495 se você estiver se sentindo particularmente arbitrário. O segundo passo é clicar com o mouse no espaço em branco abaixo, próximo de onde diz *Value label* (Valor do rótulo) (ou pressionar *Alt* e ao mesmo tempo *e*) e digite um rótulo apropriado para o grupo. Na figura 2.4, eu digitei 2 como meu código e atribuí um rótulo de *Student* (Estudante). O terceiro passo é adicionar esse código na lista clicando em Add. Na Figura 2.4, eu já defini o código 1 para o grupo dos professores; para acrescentar o código do grupo dos estudantes clique em Add. Quando você definir todos os valores dos códigos, clique em OK; se você clicou no OK e esqueceu de adicionar o último código da lista, o SPSS irá exibir uma mensagem avisando que quaisquer alterações pendentes serão perdidas. Em linguagem simples ele está dizendo para voltar e clicar no Add antes de continuar. Finalmente variáveis de código sempre representam categorias e assim os níveis em que elas são mensuradas é nominal (ou ordinal se as categoriais tiverem uma ordem), assim é necessário especificar o nível em que a variável foi mensurada indo até a coluna denominada *Measure* (Medida) e selecionando *Nominal* (ou *Ordinal* se os grupos tiverem uma ordem) de uma lista suspensa.

Tendo definido os códigos, você pode então mudar para o painel editor de dados e digitar os valores numéricos na coluna apropriada (se uma pessoa for um professor, digite 1, se for um estudante digite 2). O que é realmente excitante é que você pode solicitar que o computador apresente os códigos ou os valores que você forneceu clicando em

Material com direitos autorais

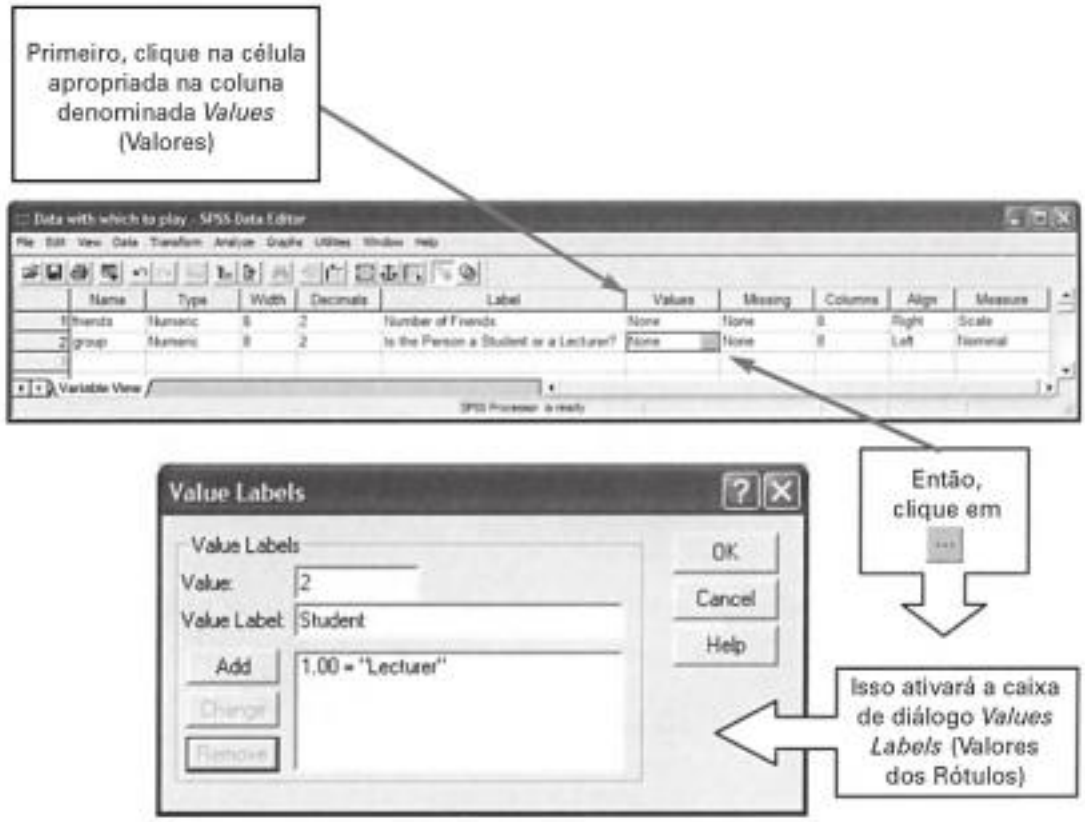


Figura 2.4 Definindo variáveis de código e seus valores no SPSS.

(veja a Figura 2.5). A Figura 2.5 mostra como os dados devem ser arranjados para uma variável de grupo/código. Agora, lembre que cada linha do editor de dados representa um participante e assim, neste exemplo, está claro que os cinco primeiros são professores enquanto que os de 6 a 10 são estudantes. Esse exemplo, também, demonstra por que variáveis de grupo são utilizadas para variáveis que tenham sido mensuradas entre assuntos: porque pela utilização do código é impossível para um participante pertencer a mais do que um grupo. Essa situação deve ocorrer no delineamento entre grupos (isto é, cada participante deve ser testado em apenas um dos grupos: o experimental ou o controle). Entretanto, no delineamento de medidas repetidas (entre conteúdos) cada participante é testado em cada uma das condições e assim nós não iremos utilizar esse tipo de variável de grupo (pois cada participante deve tomar parte em cada uma das condições experimentais).

2.4.5 Tipos de variáveis ①

Existem diferentes tipos de variáveis que podem ser usadas no SPSS. Na maioria dos casos você irá utilizar variáveis **numéricas**. Essas variáveis são aquelas que contêm números e incluem as do tipo código que acabaram de ser descritas. No entanto, uma das outras opções ao se criar uma variável é especificar o seu tipo, o que é feito clicando na coluna denominada Type no painel variável e então clicando em ... (exatamente como foi feito na especificação dos rótulos na Figura 2.4). Isso irá ativar a caixa de diálogos *Variable type* (tipo de variável) como na Figura 2.6, que mostra os *Default* (valores por omissão). Por padrão, uma variável é estabelecida como sendo numérica e armazena oito dígitos, mas você pode trocar esse valor digitando um novo em um espaço denominado de *Width* (tamanho) na caixa de diálogos. Sob circunstâncias normais você não necessitará mais do que o SPSS retenha mais

Material com direitos autorais

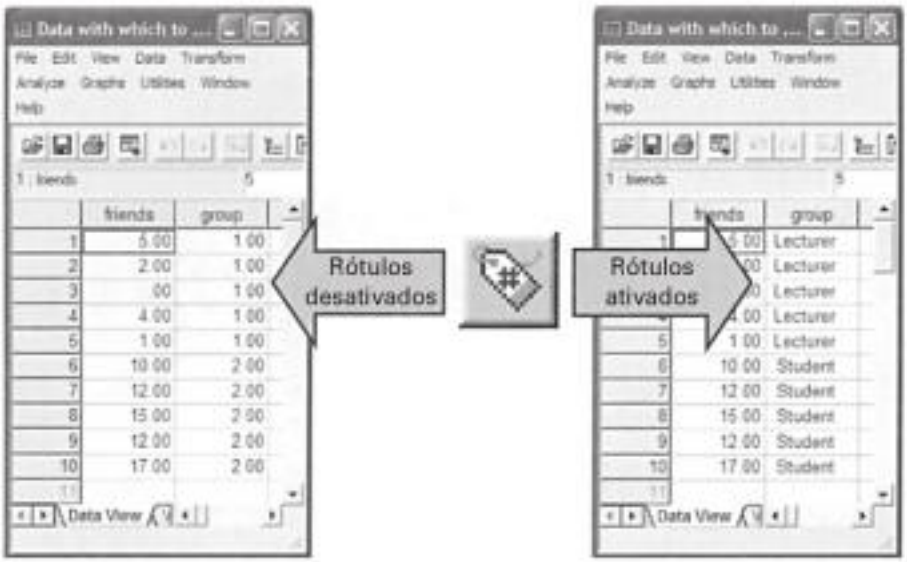


Figura 2.5 Codificando valores no editor de dados com os rótulos dos valores ativados e desativados.

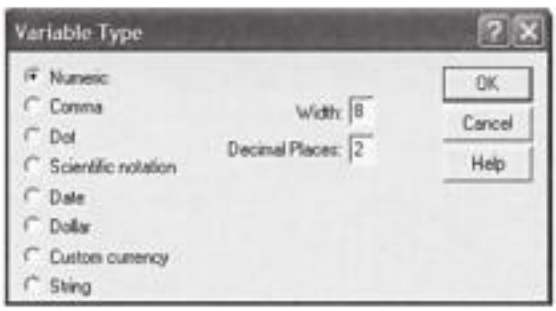


Figura 2.6 Definindo o tipo de variável sendo utilizado.

do que oito dígitos a menos que você esteja fazendo cálculos que sejam particularmente precisos. Outro valor determinado por omissão ou padrão são as duas casas após a vírgula (de fato você irá notar que por padrão sempre que você digitar qualquer número inteiro na planilha o SPSS acrescentará uma vírgula seguida de dois zeros, isso poderá ser desconcertante de início!) É bastante fácil mudar o número de casas decimais para uma dada variável substituindo o 2 por um novo valor que dependerá do grau de precisão que for necessário.

O *variable type* (tipo da variável) da caixa de diálogos também permite que se especifique outros tipos de variáveis, na maior parte serão utilizados valores numéricos. Outro tipo de variável é a texto (*string*). Uma variável texto é simplesmente uma linha de caracteres e pode representar comentários sobre um determinada participante ou outra informação que não se deseja analisar como uma variável de agrupamento (como o nome do participante). Se for selecionada a opção variável texto (*string*), o SPSS permitirá que se especifique o tamanho da variável texto (*string*), que por omissão é de oito caracteres, de forma que será possível inserir frases longas se necessário.

2.4.6 Valores que faltam (*missing*) ①

Apesar de lutarmos para coletar conjuntos de dados completos ocorre que, muitas vezes, nós temos dados que não podem, por algum motivo, ser obtidos, dados que faltam ou são desconhecidos. A omissão de dados pode ocorrer por vários motivos: em questionários longos os participantes podem acidentalmente esquecer-se de responder a algumas questões; em procedimentos experimentais falhas mecânicas podem fazer que o dado não seja registrado e em levantamento de assuntos delicados (por exemplo, comportamento sexual) os participantes podem exercer o seu direito de

Material com direitos autorais

não responder as questões. No entanto, só porque não temos informações sobre alguns detalhes dos participantes isso não significa que devemos ignorar os demais dados que temos (embora isso possa em algumas vezes criar dificuldades estatísticas). Nesses casos precisamos informar ao computador que o dado de algum participante em particular está faltando. O princípio por trás dos valores desconhecidos é totalmente similar aquele sobre variáveis de grupo ou codificadoras, isto é, aqui devemos escolher ou arbitrar um valor para representar o ponto de dados que está faltando. Esse valor simplesmente diz ao computador que não existe um valor registrado para o participante para uma determinada variável. O computador então irá ignorar a célula do editor de dados (ele não utiliza o valor em uma análise). Você deve ter cuidado na escolha desse valor e certificar-se que ele não irá coincidir com qualquer valor existente no conjunto de dados original. Por exemplo, se informarmos ao computador que o valor 9 indica um valor desconhecido ou omitido e vários participantes apresentaram genuinamente esse escore, então o computador irá tratar esses dados como não-existentes, quando na realidade, eles existem.

Para especificar um código para os valores desconhecidos basta clicar na coluna denominada de Missing no painel das variáveis do editor e então clicar em ... para ativar a caixa de diálogos *Missing Values* (valores que faltam) na Figura 2.7. Por omissão, o SPSS assume que não existem valores desconhecidos mas se você, de fato, tiver dados com esse tipo de problema você pode escolher defini-los de três formas: a primeira é selecionar valores isolados (clicando no círculo próximo onde diz *Discrete missing values* – valores desconhecidos islolados) que são valores únicos que representam dados que faltam. A razão pela qual você pode querer escolher ter vários números para representar valores que faltam é que você poderá atribuir diferentes significados para cada um dos valores. Por exemplo, você poderá ter o número 8 representando a resposta 'não se aplica', um código 9 para representar uma resposta 'eu não sei' e um código 99 dizendo que o participante não respondeu a questão. Para o computador não fará diferença e ele irá ignorar todas as células que contém esses valores; entretanto, a utilização de códigos diferentes pode ser uma forma útil de relembrar porque um determinado valor está faltando. Normalmente, um valor discreto é suficiente e em um experimento em que atitudes são mensuradas em uma escala de 100 pontos (escores variando de 1 a 100) você pode escolher o valor 666 para representar valores que faltam porque (1) esse valor não irá ocorrer nos dados que foram coletados (2) porque valores que faltam criam problemas estatísticos e você poderá lembrar que os que não responderam a questões do questionário são filhotes do capeta! Uma segunda opção é selecionar um intervalo de valores para representar dados que faltam e isto é útil em situações na qual pode ser necessário excluir dados que tem valores entre dois pontos. Assim, podemos querer excluir todos os escores entre 5 e 10. A opção final é ter um intervalo de valores e não valores isolados.

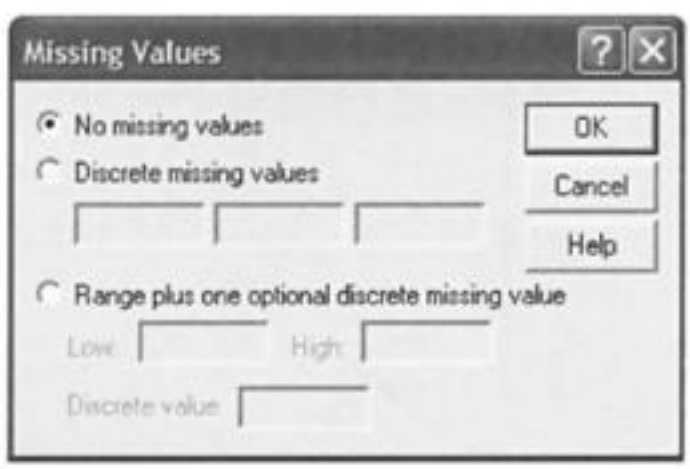


Figura 2.7 Definindo valores que faltam (*missing*).

2.4.7 Mudando o formato da coluna ①

A última opção disponível quando definimos uma variável é ajustar o formato da coluna dentro do editor de dados. A opção por omissão é ter uma coluna que tem oito caracteres de largura com todos os valores alinhas à direita. Esses dois valores podem ser alterados. No painel variáveis você pode aumentar ou diminuir o valor padrão de 8 clicando nas setas abaixo e acima ( ). Para mudar o alinhamento da coluna

apresentadas). Minha previsão na última edição desse livro foi a de que futuras versões do SPSS iriam incluir facilidades para fazer chá na janela de saída, mas infelizmente isso não se realizou (SPSS por favor tome nota!).

A Figura 2.9 mostra a estrutura básica da janela de saídas (ou de apresentação de resultados). No lado direito existe um grande espaço na qual a saída é apresentada. O SPSS mostra tanto gráficos quanto resultados de análises estatísticas nessa parte da janela. É também possível editar os gráficos e para fazer isso você simplesmente deve dar um duplo clique naquele que deseja ver editado. No lado esquerdo da janela de saída existe um diagrama de árvore ilustrando a estrutura da saída. Esse diagrama é útil para quando for realizada uma série de análises porque ela fornece uma forma fácil de acessar partes específicas da saída. A estrutura de árvore é praticamente autoexplicativa no sentido de que cada vez que for realizado um procedimento (tal como construir um gráfico ou rodar uma análise estatística), o SPSS lista esse procedimento como um cabeçalho principal.

Na Figura 2.9, construí um gráfico e então conduzi uma análise univariada da variância (ANOVA) e assim esses dois nomes aparecem como cabeçalhos principais. Para cada procedimento existe uma série de subprocedimentos e eles são listados como ramos de uma árvore debaixo do cabeçalho principal. Por exemplo,

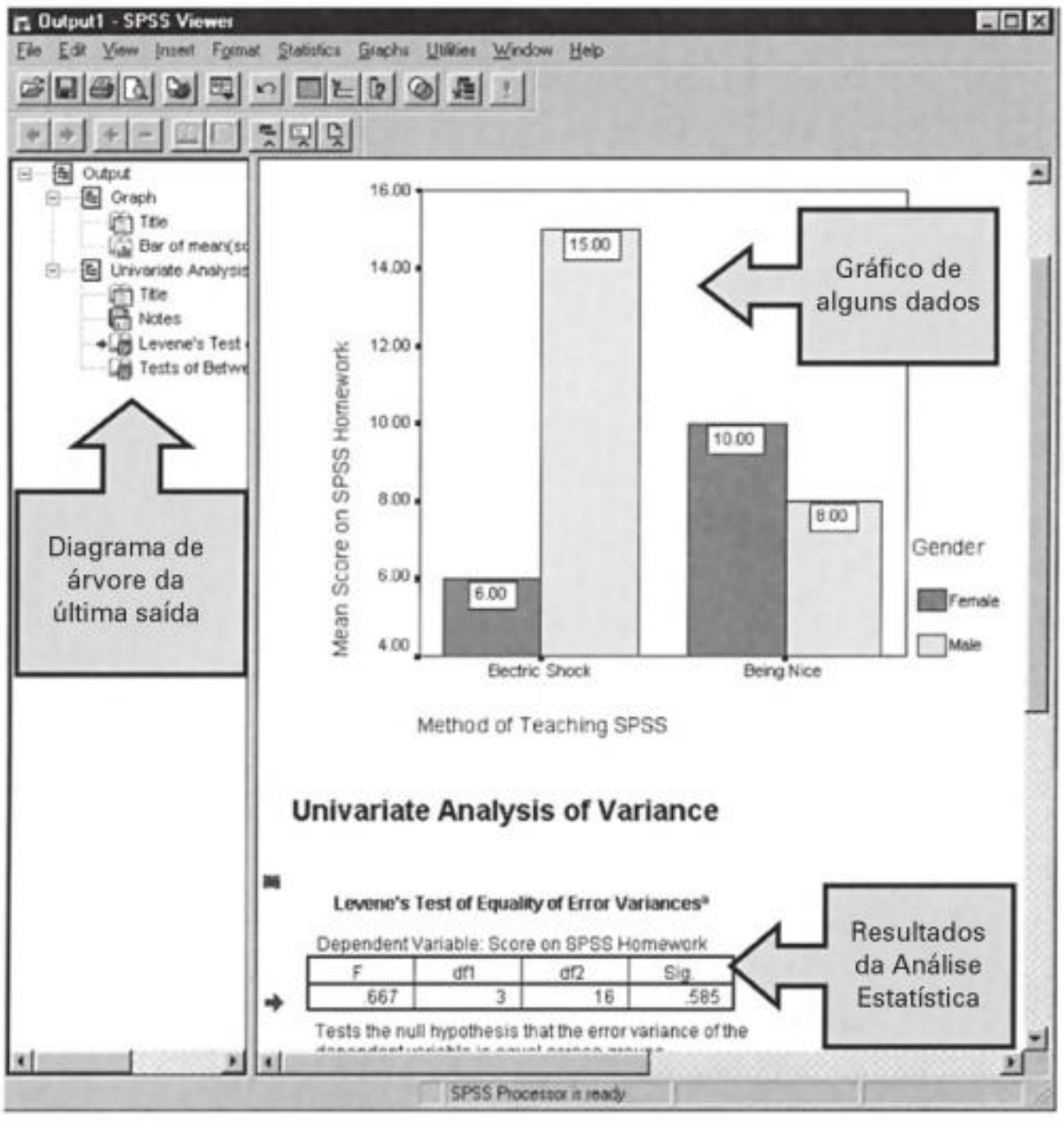


Figura 2.9 O visualizador de saídas.

uma condensação da estrutura de árvore. Isso poderá ser útil quando existir muitas saídas e o diagrama de árvore tornar-se muito complicado.

Esse ícone expande ou oculta seções. Por omissão todos os títulos principais são mostrados no diagrama de árvore em sua forma expandida. Se, no entanto, você optou por ocultar parte do diagrama (utilizando o ícone acima) então você poode utilizar esse ícone para desfazer a sua besteira.

Esse ícone e o seguinte permitem que se mostre ou oculte partes da própria saída. Assim, pode-se selecionar parte da saída no diagrama de árvore e então utilizar esse ícone para que a própria saída seja ocultada. Ela não é apagada, apenas ficará fora de vista. Assim, esse ícone é semelhante ao que oculta a estrutura de árvores só que ele afeta a própria saída em vez estrutura de árvore. Ele é útil para eliminar detalhes irrelevantes.

Esse ícone desfaz o que foi feito pelo anterior, assim se você ocultou parte da saída e clicar esse ícone ele fará essa parte oculta reaparecer. Por *Default* (omissão) todas as partes da saída são exibidas e assim esse ícone não está ativo: ele se tornará ativo somente quando você tiver ocultado parte da saída.

Embora esse ícone pareça com um rolo de pintura ele infelizmente não pinta a casa para você. O que faz é inserir um novo título ao diagrama de árvore. Por exemplo, se você realizou vários testes estatísticos relacionados a uma ou mais questões de pesquisa você poderá inserir um título principal e então rebaixar os títulos das análises relevantes de modo que todos fiquem sob esse novo título.

Considerando que você tenha feito o colocado acima, você poderá utilizar esse ícone para fornecer ao novo título uma legenda. O que você digitar irá aparecer na saída. Assim, você poderá querer ter um cabeçalho como 'Questão de Pesquisa número 1" que informará que as análise sob esse cabeçalho se relacionam com a primeira questão de pesquisa.

Esse último ícone é utilizado para colocar uma caixa de texto na janela de saída. Você poderá digitar qualquer coisa nessa caixa. No contexto dos dois ícones anteriores você poderá utilizar um quadro de texto para explicar sua primeira questão de pesquisa (por exemplo, "Minha primeira questão de pesquisa é saber se o tédio se instalou ou não ao final do primeiro capítulo do meu livro. A análise seguinte testa a hipótese de que os níveis de tédio serão significativamente mais altos ao final do capítulo do que no início.").

2.6 A JANELA DE SINTAXE ③

Eu mencionei anteriormente que algumas vezes seria útil utilizar a sintaxe do SPSS. Ela é uma linguagem de comandos para executar análises estatísticas e manipulação de dados. Na maioria das vezes nós faremos as coisa que precisamos utilizando as caixas de diálogos; entretanto, a sintaxe do SPSS poderá ser útil. Em algumas situações existirão coisas que só poderão ser executadas com o recurso da sintaxe (admito que muitas dessas coisas são bastante avançadas, mas existirão umas poucas situações nesse livro em que mostrarei a você alguns truques legais utilizando a sintaxe). Uma segunda razão para utilizarmos a sintaxe é se análises semelhantes forem executadas muitas vezes. Nessas situações é frequentemente mais rápido fazer a análise e salvar a sintaxe à medida que se prossegue. Felizmente isso é fácil de fazer porque muitas das caixas de diálogos do SPSS apresentam um botão Paste. Quando você tiver especificado a análise utilizando uma caixa de diálogos, se clicar nesse botão colorá a sintaxe utilizada em uma janela de sintaxe. Para abrir a janela de sintaxe utilize os menus **File⇒New⇒Sintax...** e uma janela de sintaxe em branco irá aparecer como a da

Material com direitos autorais

Figura 2.10. Nessa janela você poderá digitar os comandos de sintaxe que desejar. As regras da sintaxe do SPSS podem ser bastante frustrantes (por exemplo, cada linha deve terminar com um ponto e se ele for esquecido uma mensagem de erro irá aparecer) e estão além dos objetivos deste livro. Entretanto, à medida que formos avançando, mostrarei algumas coisas que darão a você um pouco do sabor de como a sintaxe poderá ser utilizada. Muitos de vocês não precisarão dela, mas para aqueles que desejarem ou precisarem espero que essa pequena amostra seja suficiente para que tenham o interesse despertado e continuem a aprender mais. Uma vez que os comandos tenham sido digitados eles precisam ser executados pela utilização do menu **Run**. O comando **Run⇒All** executará toda a sintaxe digitada na janela (clicando em ▶ isso também ocorrerá), ou você poderá marcar uma seleção da sua sintaxe com o mouse e utilizar o **Run⇒Selection** para processar a sintaxe selecionada. Você poderá rodar o comando corrente utilizando **Run⇒Current** (ou pressionar *Ctrl* e *R* no teclado) ou finalmente executar toda a sintaxe a partir do cursor até o final da janela de sintaxe utilizando **Run⇒To End**.

2.7 SALVANDO ARQUIVOS ①

Embora muitos de vocês estejam familiarizados com a tarefa de salvar arquivos no Windows, ela é um conhecimento vital e assim descreverei brevemente o que é preciso fazer. Para salvar um arquivo simplesmente utilize o ícone 💾 (ou use os menus **File⇒Save** ou **File⇒Save As...**). Se o arquivo for novo, então clicar nesse ícone ativará a caixa de diálogos *Save Data As* (Salvar Dados como) (veja a Figura 2.11). Se você estiver no editor de dados quando selecionar *Save As...* (Salvar Como) o SPSS irá salvar o arquivo em que você estiver trabalhando no momento, mas se estiver no painel de visualização de saídas ele irá salvar a saída existente no momento.

Existem várias características na caixa de diálogos da Figura 2.11. Primeiro você precisar determinar um local onde armazenar (salvar) o arquivo. Normalmente, existem dois tipos de locais onde se pode guardar dados: o disco rígido e as mídias móveis (disquete, CDs, DVDs, *Pen-Drives*, etc.) ou outras opções que nem imagino que você possa ter no seu computador. A primeira coisa a fazer é selecionar ou uma mídia móvel ou o disco rígido e dar um duplo clique em 💽. Uma vez que tenha escolhido o local principal a caixa de diálogos mostrará todas as pastas disponíveis naquela mídia em particular (você pode ainda não ter pastas no disco ou mídia e nesse caso poderá criar uma clicando em 📁). Logo que tiver selecionado uma pasta na qual irá salvar o arquivo você precisará dar-lhe um nome. Se você clicar no espaço próximo onde diz *File name* (Nome do Arquivo) um cursor irá aparecer e você poderá digitar um nome de até dez letras. Por omissão o arquivo será salvo no formato do SPSS, assim ele apresentará a extensão *.sav*, se ele for um documento do visualizador de saídas ou janela de saída o arquivo terá a ex-

Figura 2.10 Uma nova janela de sintaxe (esquerda) e uma com alguma sintaxe (direita).

Material com direitos autorais

Figura 2.11 A caixa de diálogos *Save Data As* (Salvar Dados Como).

tensão *.spo* e se ele for um arquivo de sintaxe a extensão será *.sps*. Entretanto, você poderá salvar dados em outros formatos tais como os do Microsoft Excel e arquivos textos delimitados por tabulações. Para fazer isso clique em onde diz *Save as type* (Salvar como tipo) e uma lista dos possíveis formatos será apresentada. Clique no tipo que você desejar. Uma vez que um arquivo tenha sido previamente salvo, ele poderá ser salvo novamente (atualizado) clicando em Esse ícone aparece tanto no editor de dados quanto no visualizador de saídas e o arquivo a ser salvo dependerá da janela que estiver ativa no momento. O arquivo será salvo no local em que já estiver armazenado.

2.8 RECUPERANDO UM ARQUIVO ①

Ao longo do livro você irá trabalhar com arquivos de dados que estão disponíveis no *site* www.artmed.com.br. Dessa forma é importante que você saiba como carregar esses arquivos de dados no SPSS. *O procedimento é bastante simples. Para abrir um arquivo simplesmente utilize o ícone (ou utilize o menu ***File⇒Open⇒Data...*** [Arquivo⇒Abrir⇒Dados]) para ativar a caixa de diálogos como na Figura 2.12. Primeiro você precisa encontrar o local onde o arquivo está armazenado (salvo). Se você está carregando o arquivo a partir do CD então acesse o diretório do CD clicando em , onde diz *Look in* (Olhar em) e uma lista de localização de diretórios será mostrada. Uma vez que o diretório do CD tenha sido acessado (clicando em) você deve ver uma lista de arquivos e pastas que podem ser abertas. Como ocorreu com o salvamento de um arquivo, você está no editor de dados então o SPSS irá mostrar apenas arquivos de dados do SPSS para serem abertos (se você estiver na janela de visualização de saída então somente arquivos de saída serão mostrados). Se você utilizar os menus e utilizar o caminho ***File⇒Open⇒Data...*** (Arquivo⇒Abrir⇒Dados) então arquivos de dados serão apresentados, mas se utilizar o caminho ***File⇒Open⇒Output...*** (Arquivo⇒Abrir⇒Saídas) então arquivos de saídas serão exibidos e se usar ***File⇒Open⇒Sintax...*** (Arquivo⇒Abrir⇒Sintaxe) então arquivos de sintaxe serão mostrados (você deve ter captado a ideia geral). Você pode abrir uma pasta dando um duplo clique sobre o seu ícone. Uma vez que você tenha chegado ao arquivo desejado você poderá abrí-lo selecionando-o

* N de R. Na edição original, os arquivos foram disponibilizados em um CD. Nesta edição, os mesmos arquivos devem ser acessados no *site* www.artmed.com.br e salvos em seu computador (conteúdo em inglês).

Material com direitos autorais

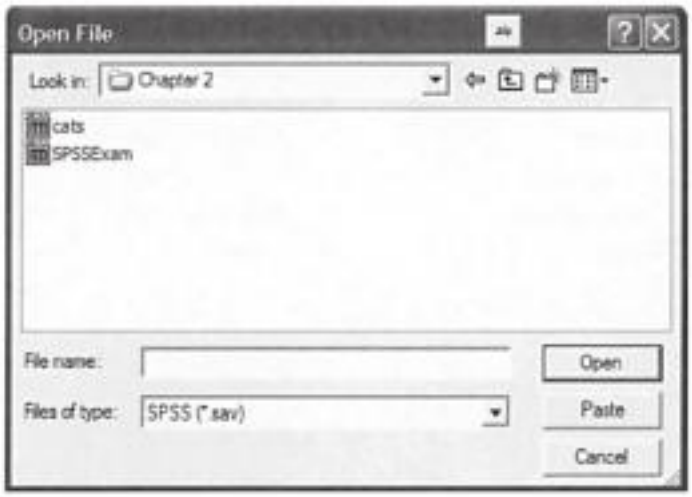


Figura 2.12 Caixa de diálogos para abrir um arquivo.

com o mouse ou então clicando no Open, ou ainda clicando duas vezes no ícone próximo ao arquivo que deseja abrir (por exemplo, clicando duas vezes em cats). Os dados do arquivo aparecerão então na janela apropriada. Se você está no editor de dados e quiser abrir um arquivo de saída, clique em ▼ onde diz *Files of type* (Arquivos do tipo) e uma lista alternativa de formatos de arquivos será apresentada. Clique no tipo de arquivo apropriado (documento de saída (**.spo*), arquivo de sintaxe (**.sps*), arquivo do Excel (**.xls*), arquivo texto (**.dat*, **.txt*) e um arquivo de desse tipo estará disponível para ser aberto.

2.9 O QUE DESCOBRIMOS SOBRE ESTATÍSTICA? ①

Este capítulo forneceu uma introdução básica ao ambiente do SPSS. Vimos que ele utiliza duas janelas principais: um editor de dados e um visualizador de saídas. O editor de dados tem dois painéis um de dados (onde entramos com os valores ou dados) e um de variáveis (onde são definidas as variáveis e suas propriedades). O visualizador é uma janela na qual toda saída aparece, tais como tabelas, estatísticas e gráficos. Nós também criamos nosso primeiro conjunto de dados definindo algumas variáveis e entrando com alguns dados. Ao fazer isso, descobrimos que podemos codificar grupos de pessoas utilizando números (variáveis codificadoras) e descobrimos que linhas no editor de dados representam pessoas (ou casos de dados) e colunas representam diferentes variáveis. Finalmente damos uma olhada na janela de sintaxe e aprendemos como abrir e salvar arquivos.

2.10 TERMOS-CHAVE QUE DESCOBRIMOS

- Dados de razão
- Dados intervalares
- Dados nominais
- Dados ordinais
- Editor de dados
- Editor de sintaxe
- Variáveis numéricas
- Variáveis texto
- Visualizador
- Visualizador de dados

2.11 TAREFA DO ALEX ESPERTO

A tarefa do Alex Esperto para este capítulo é salvar os dados que foram digitados. Salve-os em algum lugar do disco rígido do seu computador (ou em uma mídia móvel se preferir). Dê um título significativo (crie, talvez, uma pasta

Material com direitos autorais

chamada "Meus arquivos de dados" em "Meus documentos" um diretório do Windows se você estiver utilizando um PC.①

2.12 LEITURAS COMPLEMENTARES

EINSPRUCH, E. L. *An Introductory guide to SPSS for Windows*. Thousand Oaks (CA): Sage, 1998.

FOSTER, J. J. *Data Analysis using SPSS for Windows: a beginner's guide*. London: Sage, 2001 (2ª ed.).

KINNEAR, P. R., GRAU, C. D. *SPSS for Windows made simple (Release 10)*. Hove: Psychology Press, 2000.

Material com direitos autorais

3
EXPLORANDO DADOS

3.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ①

Como o título sugere, este capítulo ensinará o primeiro passo da análise de dados: explorá-los. Wright (2003), citando Rosenthal, afirma que pesquisadores devem "ficar amigos dos seus dados". Não, isso não significa que quem utiliza estatística deve ser amigo dos seus dados porque são os únicos amigos que eles têm; em vez disso, Rosenthal quer dizer que os pesquisadores muitas vezes fazem análises apressadas. Wright faz a analogia com um bom vinho; você deve saborear o aroma e os sabores delicados para realmente apreciar a experiência. Isso talvez exagere a beleza da análise de dados; no entanto, apressar a análise é, suponho, parecido com engolir o vinho: o resultado é confuso e incoerente! Este capítulo apresenta algumas formas de exploração de dados. Começaremos verificando como podemos examinar os dados, checar algumas hipóteses básicas sobre eles e então determinar algumas estatísticas descritivas e traçar alguns gráficos.

3.2 DADOS PARAMÉTRICOS ①

3.2.1 Hipóteses "retirar" ①

Muitos dos procedimentos estatísticos descritos neste livro são **testes paramétricos** baseados na distribuição normal (descrita na Seção 1.5). Um teste paramétrico requer que os dados sejam retirados de um grande catálogo de distribuições descritas por estatísticos e para as quais certas suposições devem ser verdadeiras. Se você utiliza um teste paramétrico quando seus dados não são paramétricos, os resultados talvez não sejam apropriados. Dessa forma, é importante que você verifique as hipóteses antes de decidir qual teste estatístico é apropriado. Ao longo deste livro você conhecerá minha obsessão com hipóteses e sua checagem. A maioria testes paramétricos baseados na distribuição normal deve preencher quatro hipóteses básicas para que seja preciso. Muitos estudantes acham que testar suposições é uma tarefa entediante e frequentemente ficam confusos sobre como saber se uma hipótese é ou não-verdadeira. Assim, este capítulo foi projetado para levá-lo passo a passo em uma turnê pelo mundo das suposições

Material com direitos autorais

paramétricas (oh, que excitante!). Até o momento, você pode achar que essas suposições não são muito atraentes, mas elas podem ter grandes vantagens: pelo menos você pode impressionar seus professores/orientadores listando todas as hipóteses dos testes que eles violaram ao longo da carreira. Então você pode esquecer, no campo da estatística, as teorias que eles levaram a vida desenvolvendo – e eles não poderão censurá-lo.[1] Bem, como sempre digo, se você precisa encarar a agonia de trabalhar com estatística, deve fazê-la de forma apropriada. As hipóteses dos testes paramétricos são:

1. **Dados normalmente distribuídos:** Assume-se que os dados foram obtidos de uma ou mais populações normais. O fundamento dos testes de hipóteses baseia-se em ter populações distribuídas normalmente. Assim, se essa hipótese não é satisfeita, a lógica que os sustenta será falha (esses princípios já mencionados no Capítulo 1 e serão explicados novamente no Capítulo 6). Muitos pesquisadores verificam suas amostras (utilizando um histograma) e se a amostra parece aproximadamente normal, eles assumem que as populações também serão. Veremos, neste capítulo, que podemos ir um passo além dessa abordagem e testar se a amostra difere significativamente de uma normal.
2. **Homogeneidade da variância:** Essa hipótese significa que as variâncias devem ser as mesmas para as diferentes populações sendo consideradas. Em delineamentos nos quais vários grupos de participantes são testados, essa hipótese significa que cada uma das amostras foi retirada de populações com a mesma variância. Em delineamentos correlacionais, essa hipótese significa que a variância de uma variável deve ser estável em relação a todos os níveis da outra variável (veja a Seção 3.6).
3. **Dados por intervalo:** Os dados devem ser mensurados pelo menos ao nível ordinal (veja o Quadro 2.1 no Capítulo 2). Isso significa que a distância entre os pontos da sua escala deve ser igual em todas as partes ao longo da escala. Por exemplo, se você tem uma escala de ansiedade de 10 pontos, a diferença de ansiedade representada por uma alteração de 2 para 3 no escore deve ser a mesma representada por uma alteração do escore de 9 para 10.
4. **Independência:** Essa hipótese considera que os dados de participantes diferentes são independentes, o que significa que o comportamento de um participante não influencia o comportamento de outro. Nos delineamentos de medidas repetidas (na qual os participantes são mensurados em mais de uma condição experimental), esperamos que os escores na condição experimental sejam dependentes para um dado participante, mas que o comportamento entre diferentes participantes deve ser independente. Como exemplo, imagine que duas pessoas, Paulo e Júlia, estão participando de um experimento onde eles devem indicar se lembram de ter visto determinadas fotos anteriormente no experimento. Se Paulo e Júlia pudessem conversar sobre se viram ou não certas fotos, então suas respostas *não* seriam independentes. A resposta de Júlia a uma determinada questão dependerá da resposta de Paulo e isso viola a hipótese de independência. Se Paulo e Júlia não pudessem conversar (se estivessem em salas diferentes), suas respostas seriam independentes (a menos que eles fossem telepatas!): a resposta de Paulo não seria influenciada pela de Júlia e vice-versa.

[1] Quando eu estava fazendo meu doutorado, recebemos a tarefa do nosso professor de estatística de pegar algum artigo publicado e criticar os métodos estatísticos utilizados. Escolhi um do meu orientador e resolvi criticar duramente todos os procedimentos de análise de dados (e eu fui *bastante* pedante). Imagine o meu horror quando vi meu orientador aproximando-se aos pulos pelo corredor com um largo sorriso e declarando que, sem que eu soubesse, ele foi o segundo avaliador do meu ensaio. Felizmente, ele teve senso de humor e eu tirei uma boa nota.☺

Material com direitos autorais

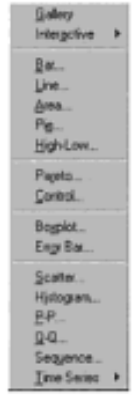
As hipóteses de dados intervalares e medidas independentes são, infelizmente, testadas somente pelo senso comum. A hipótese da homogeneidade da variância é testada de forma diferente por diferentes procedimentos e, assim, embora abordemos esta hipótese na Seção 3.6, também iremos discutir esses testes sempre que for necessário. Isso nos deixa somente com a hipótese de normalidade para checar e em muitos aspectos essa é, de qualquer forma, a hipótese mais importante. A maneira mais simples de testar essa hipótese é olhar para a distribuição da amostra. Se os dados da amostra forem normalmente distribuídos, nossa tendência é assumir que eles vieram de uma população normal. Da mesma forma, se a nossa amostra não for normalmente distribuída, iremos assumir que ela veio de uma distribuição não-normal. Assim, para testar essa hipótese, podemos começar plotando com que frequência cada escore ocorre (*a distribuição de frequências*) e ver se ela parece normal. Vamos ver como se faz isso.

3.3 APRESENTANDO DADOS GRAFICAMENTE ①

Vamos dar uma olhada em como podemos utilizar uma distribuição de frequências, outros gráficos e estatística descritiva para fazer uma análise completa de nossos dados. Vamos usar um exemplo para ilustrar o que queremos fazer. Uma bióloga está preocupada com os efeitos potenciais dos festivais de música sobre a saúde. Assim, em determinado ano ela foi ao Festival de Música de Glastonbury (os não-britânicos podem fingir que o festival é o Roskilde, Ozzfest, Lollapalooza, Rock in Rio ou outro qualquer!) e verificou a higiene de 810 fãs durante os três dias de duração do festival. Em teoria, cada pessoa foi avaliada diariamente, mas em virtude da dificuldade de fazer isso, faltaram alguns dados nos dias dois e três. A higiene foi mensurada com a utilização de uma técnica padrão (não se preocupe: não foi lamber a axila da pessoa) que resultou em um escore variando de 0 (você está cheirando como um cadáver em decomposição em cima de um gambá) a 5 (você cheira a rosas em um dia de primavera). Hoje, eu sei, por amargas experiências, que a higiene nem sempre é o forte nesses lugares (o Festival de Reading parece ser particularmente ruim…) e, assim, essa pesquisadora previu que a higiene pessoal diminuiria muito durante os três dias do festival. O arquivo de dados denominado **GlastonburyFestival.sav** pode ser encontrado no *site* www.artmed.com.br (veja a Seção 2.8 para lembrar como abrir um arquivo)

3.3.1 Passo 1: Apontando os erros óbvios utilizando histogramas ①

A primeira coisa a fazer é construir o histograma para ver a forma dos dados. O SPSS tem recursos para construir diferentes tipos de gráficos. Se você clicar no menu **Graphs** (Gráficos), uma lista aparecerá aparecer mostrando todas as opções disponíveis. Você já deve conhecer algumas delas (por exemplo, diagramas de *pizza*, barras, etc.):



- **Bar** (Barras)**:** Normalmente utilizado para representar médias de grupos diferentes de pessoas (*Summaries for groups of cases* – resumos para grupos de casos) ou média de variáveis diferentes (*Summaries of separate variables* – resumos de diferentes variáveis).
- **Line** (Linha)**:** Também utilizado para representar médias.
- **Pie** (Torta)**:** Utilizado para apresentar frequências e percentagens.
- **Boxplot** (Caixa e Bigode)**:** Utilizado para mostrar a mediana, espalhamento e intervalo interquartílico dos escores.
- **Error Bar** (Barra de Erros)**:** Mostra a média e o intervalo de 95% de confiança em torno dessa média (veja o Capítulo 7).
- **Scatter** (Dispersão)**:** Mostra relações entre duas variáveis (veja o Capítulo 5).

Material com direitos autorais

- **Histogram** (Histograma): Mostra a frequência de diferentes escores (útil para visualizar melhor a distribuição dos escores)

Você irá notar que existe também uma opção denominada *Interactive* (Interativa), que abre outra lista com praticamente os mesmos gráficos desse menu. Os gráficos interativos são na verdade esteticamente mais agradáveis e supostamente devem mudar quando os dados forem alterados (por isso o nome!). Assim, para fazer um histograma você deve selecionar **Graphs⇒Interactive⇒Histogram...** (Gráficos⇒Interativo⇒Histograma) a fim de abrir uma caixa de diálogo como a da Figura 3.1.

Note que as variáveis do editor de dados estão listadas no painel da esquerda e você pode arrastar cada uma delas para qualquer um dos espaços em branco. Os dois eixos são representados por setas e você simplesmente arrasta a variável para o espaço adequado. No eixo vertical já existe uma variável denominada *Count* (Contagem). Isso, porque queremos contar a frequência de cada escore e o SPSS fará isso. Para começar, vamos fazer o histograma dos dados da higiene para o primeiro dia do festival. Selecione essa variável da lista e arraste para o espaço do eixo horizontal. A caixa de diálogo final deve ser parecida com a da Figura 3.1. Você notará um espaço denominado *Panel variables* (Painel de variáveis) e, se você arrastar a variável para o espaço, o SPSS irá criar um histograma para cada nível dessa variável. Assim, se você tiver uma variável de grupo (codificadora) para gênero e você a colocar no espaço, obterá dois histogramas: um para as mulheres e outro para os homens. As opções padrão (por omissão) são adequadas e, dessa forma, você pode apenas clicar no OK (se quiser ajustar uma curva normal ao histograma resultante, deverá selecionar na guia *Histogram* (Histograma) a opção *Normal curve* (Curva normal)).

O histograma resultante é mostrado na Figura 3.2 e você logo notará que existe um caso que é bem diferente dos demais. Todos os escores parecem estar concentrados em um lado da distribuição porque eles são menores do que 5 (formando uma distribuição leptocúrtica) com exceção de um, que apresentou um valor de 20. Ele é um valor atípico (*outlier*): um valor bem diferente dos demais. Valores atípicos distorcem a média e inflacionam o desvio padrão (veja Field e Hole, 2003) e uma análise criteriosa dos dados é uma forma de detectá-los. O estranho desse valor é que ele está acima do valor máximo da escala utilizada (lembre-se: a escala de higiene varia de 0 a 5) e ele deve ser um erro (ou a pessoa tem um transtorno obsessivo-compulsivo e se esfregou tanto que

Figura 3.1 Caixa de diálogo para um histograma interativo (mudar as guias no topo da caixa de diálogo altera o conteúdo da mesma).

Material com direitos autorais

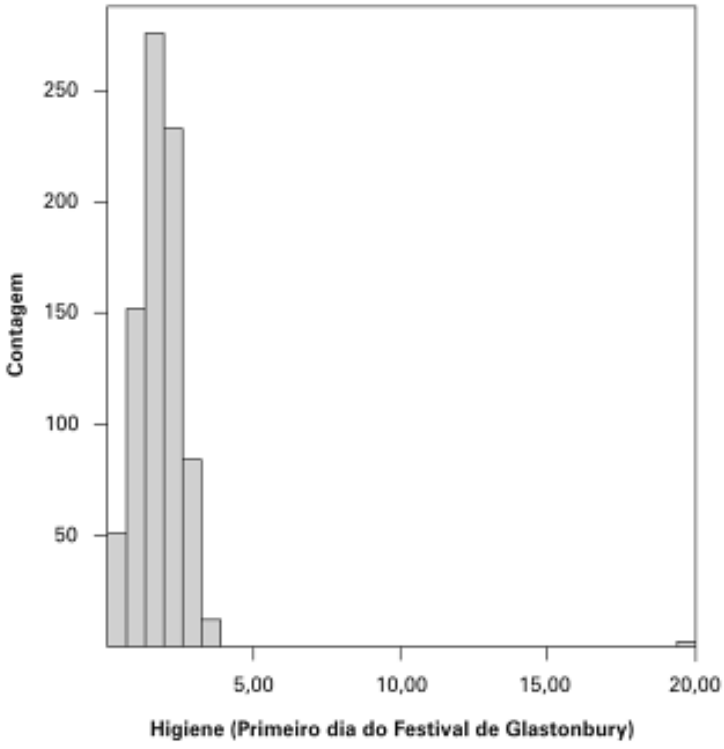


Figura 3.2 Histograma dos escores da higiene (dia 1) do Festival de Glastonbury.

atingiu um estado de limpeza extrema). Entretanto, com 810 casos, como iremos encontrar qual caso foi digitado errado? Podemos vasculhar os dados, mas certamente teremos uma dor de cabeça; em vez disso, podemos traçar um diagrama de **caixa e bigodes**.

Falarei mais sobre diagramas de caixa e bigodes posteriormente neste capítulo, assim, por ora, apenas acredite em mim e selecione **Graphs⇒Boxplot** (Gráficos⇒Diagrama de Caixa e Bigodes) para acessar a caixa de diálogo. Dentro da janela, selecione *Summaries of separate variables* (Diagrama para variáveis separadas) e clique em Define (Definir). Isso abrirá outra caixa de diálogo denominada *Boxes Represent* (Os quadros representam). Selecione as três variáveis (representando os três dias do festival) da lista e transfira todas para o painel denominado *Boxes Represent* (Os quadros representam) clicando em ▶. Agora, clique em Options (Opções) e, na caixa que abrir, selecione *Exclude cases variable by variable* (Exclua casos variável por variável). Precisamos selecionar essa opção porque em algumas das nossas variáveis faltam valores e, se excluirmos os casos por listagem (*Exclude cases listwise*), então se um caso tem um único valor omitido em qualquer uma das três variáveis, ele não será utilizado. Obviamente, não queremos isso; queremos excluir somente os casos para a variável na qual aquele valor esteja faltando. Para tanto, basta selecionar a opção *Exclude cases variable by variable* (excluir casos variável por variável). A caixa de diálogo final será semelhante à Figura 3.3. Clique em OK para obter os gráficos.

O diagrama de caixa e bigodes resultante é apresentado na Figura 3.4. É importante notar que o que foi detectado no histograma é mostrado acima do diagrama como um asterisco com um número ao lado. Esse número é a posição do caso que produziu o valor atípico. Ele é

Material com direitos autorais

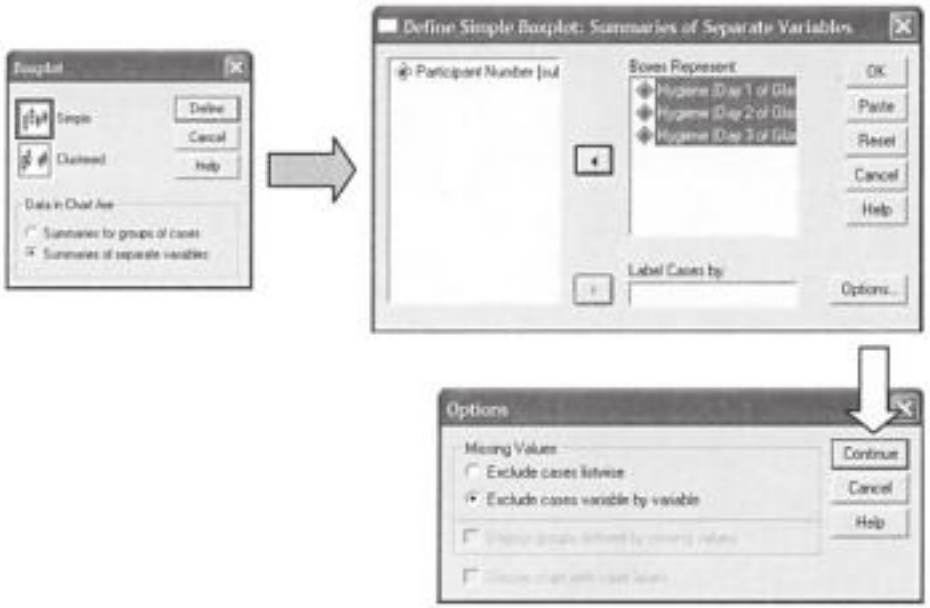


Figura 3.3 Caixa de diálogo para definir um diagrama de caixa e bigodes para diversas variáveis.

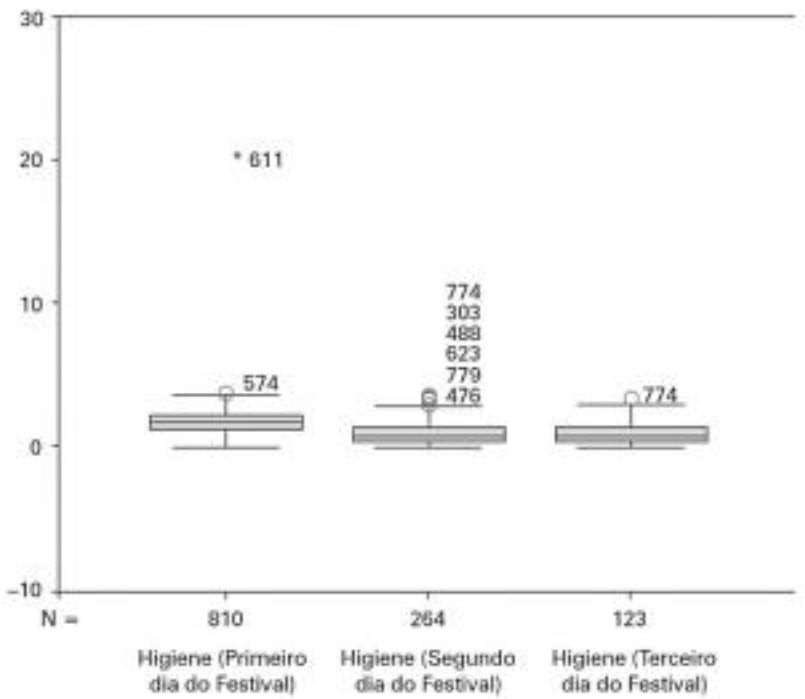


Figura 3.4 Diagrama de caixa e bigodes dos escores da higiene nos três dias do Festival de Glastonbury.

o caso de número 611. Se você for para o editor de dados poderá localizar esse caso rapidamente clicando em e digitando 611 na caixa de diálogo que aparece. Isso levará você direto para o caso 611. Olhando para esse caso, você verifica que o valor é 20,02, provavelmente um erro de digitação de 2,02. Você poderá consultar as anotações (questionários ou outro instru-

Material com direitos autorais

mento de coleta) para verificar. Vamos assumir que você fez isso e que o valor é de fato 2,02, assim, substitua o valor 20,02 por 2,02 antes de continuar com esse exemplo.

Agora que corrigimos o erro, vamos refazer o histograma. Enquanto fazemos isso devemos obter os histogramas para os dados dos dias dois e três do festival também. A Figura 3.5 mostra os três histogramas resultantes. A primeira coisa a notar é que os dados do primeiro dia parecem mais saudáveis agora que removemos o que foi digitado errado. De fato, a distribuição tem uma aparência surpreendentemente normal: ela é simétrica e não parece nem pontiaguda nem achatada – isso é bom! No entanto, o mesmo não acontece com as distribuições dos dias dois e três, que não parecem simétricas. Ambas são positivamente assimétricas. Em geral, parece que os dados dos dias dois e três estão mais concentrados em torno dos valores mais baixos da escala. Lembre que quanto mais baixo o escore, menos higiênica a pessoa é, dessa forma, isso sugere que as pessoas estão mais sujas à medida que o festival progride. A assimetria acontece porque uma minoria insiste em manter seus níveis de higiene (contra todas as chances!) ao longo do festival (lenços umedecidos são indispensáveis, eu acho). No entan-

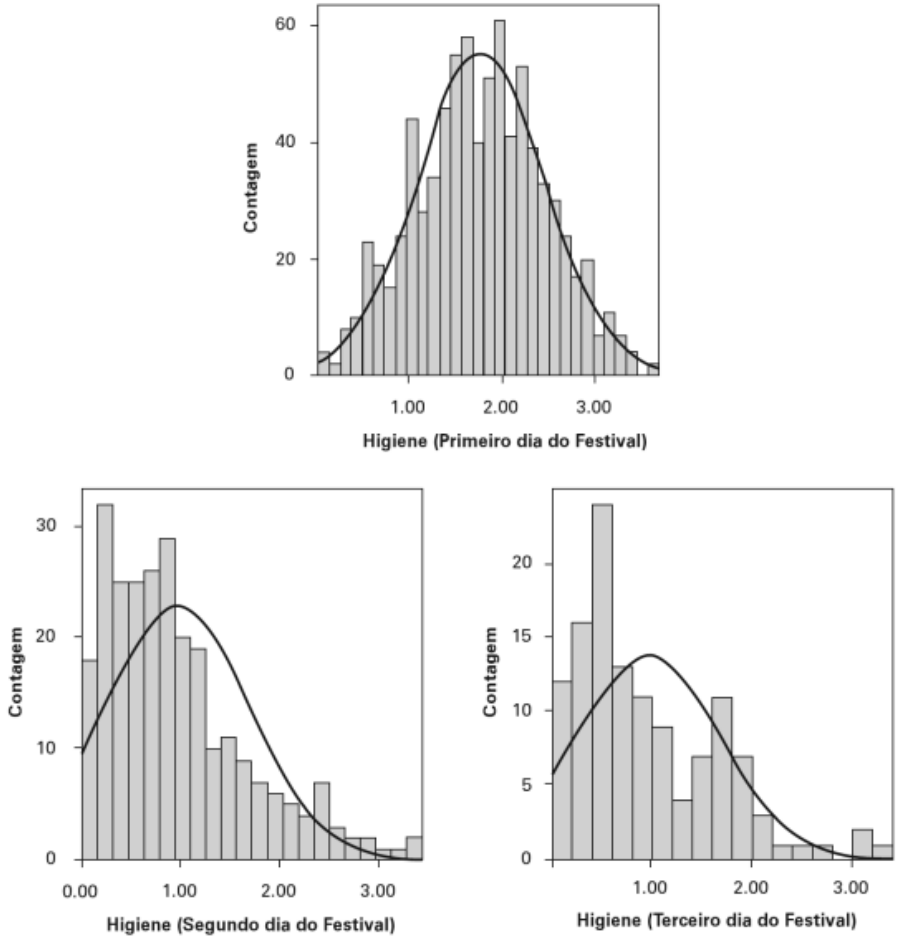


Figura 3.5 Histograma dos escores dos três dias do festival de higiene de Glastonbury.

Material com direitos autorais

to, essas distribuições assimétricas podem nos causar problemas se quisermos utilizar testes paramétricos. Na próxima seção, iremos verificar como medir a assimetria e a curtose (achatamento) dessas distribuições.

3.3.2 Passo 2: Estatísticas descritivas e o diagrama de caixa e bigodes ①

Tendo analisado rapidamente as distribuições dos escores da higiene e detectado e corrigido um valor digitado incorretamente, devemos prosseguir procurando uma maneira de quantificar a forma das distribuições e detectar valores atípicos. Para avançar na exploração das distribuições das variáveis, podemos utilizar o comando *frequencies* (frequências), que é acessado utilizando o caminho **Analyze⇒Descriptive Statistics⇒Frequencies...**[2] (Analisar⇒Estatística Descritiva⇒Frequências...). A caixa de diálogo principal é apresentada na Figura 3.6. As variáveis no editor de dados devem estar listadas no painel da esquerda e elas podem ser transferidas para o quadro denominado *Variable(s)* (Variável[is]) clicando na variável (ou destacando diversas com o mouse) e depois clicando em ▶. Qualquer análise escolhida será realizada em todas as variáveis listadas no quadro *Variable(s)* (Variável[is]). Se uma variável listada no quadro *Variable(s)* (Variável[is]) é selecionada com o mouse, ela pode se transferida de volta para a lista de variáveis clicando no botão com a seta (que deve agora estar apontando para a direção oposta). Por padrão, o SPSS produz uma distribuição de freqüências de todos os escores no formato de tabela. No entanto, existem duas outras caixas de diálogo que podem ser utilizadas para produzir outras opções. A caixa de diálogo *Statistics* (Estatísticas) é acessada clicando em Statistics... e a caixa de diálogo *charts* (gráficos) pode ser ativada clicando em Charts...

[2] Lembre que esse caminho de menu será **Statistics⇒Summarize⇒Frequencies...** (Estatística⇒Resumir⇒Frequências) na versão 8.0 e anteriores.

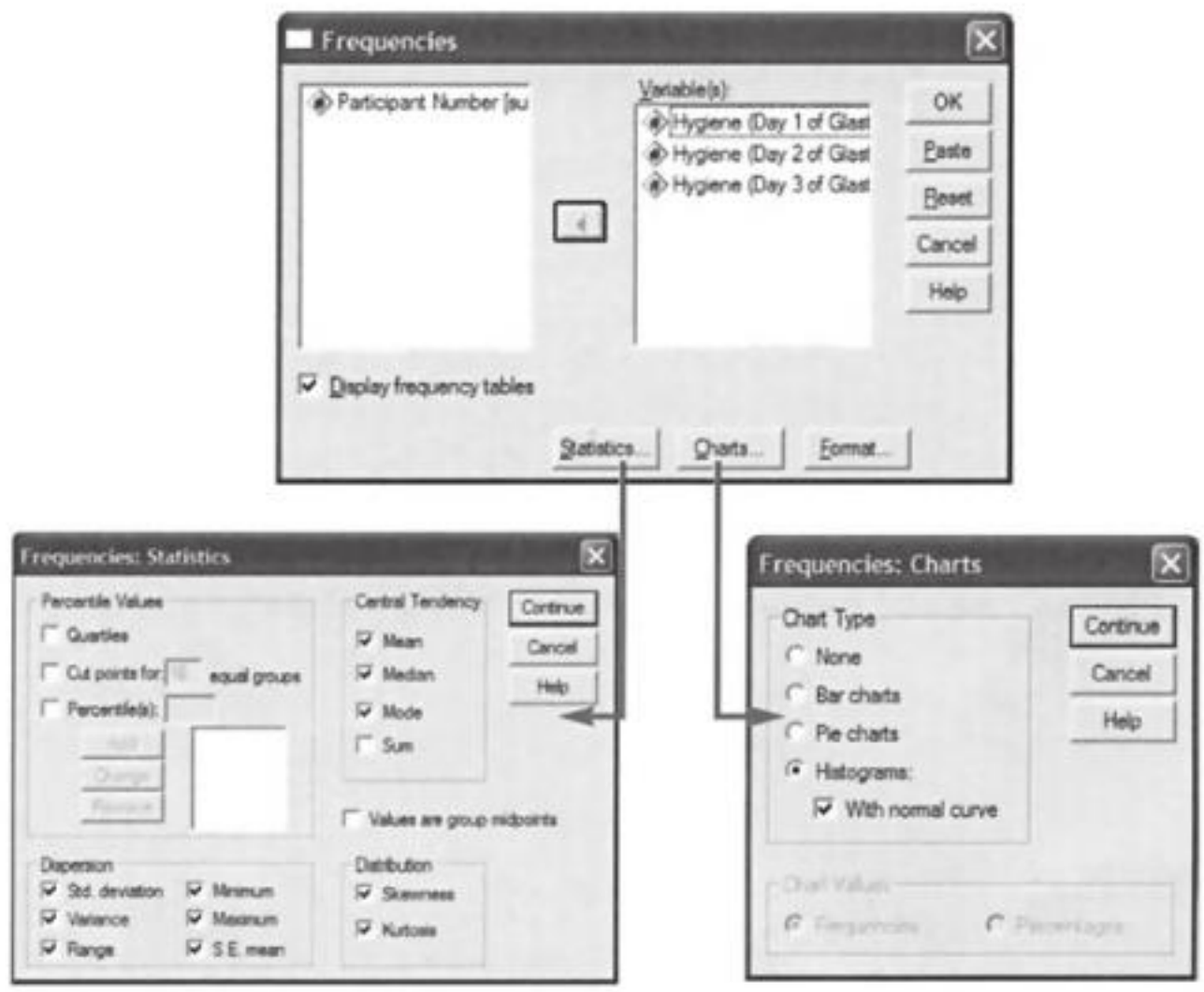


Figura 3.6 Caixa de diálogo para o comando *frequencies* (frequências).

A caixa de diálogo *Statistics* (Estatística) permite a seleção de várias opções de maneiras como a distribuição dos escores pode ser descrita, como medidas de tendência central (média, moda, mediana), medidas de variabilidade (amplitude, desvio padrão, variância, quartis), medidas de forma (assimetria e curtose). Para descrever as características dos dados, devemos selecionar a média, moda, mediana, desvio padrão, variância e amplitude. Para verificar se uma distribuição é normal, precisamos olhar os valores da assimetria e da curtose. A opção *chart* (diagramas) fornece uma maneira simples de representar graficamente a distribuição dos escores (como um gráfico de barras, um de *pizza* ou um histograma). Já obtivemos histogramas dos nossos dados, assim, não precisamos selecionar essa opção, mas poderemos utilizá-la em análises futuras. Quando você tiver selecionado as opções adequadas, retorne à caixa de diálogo principal clicando em Continue. Uma vez lá, clique em OK para obter a análise.

A saída 3.1 do SPSS mostra a tabela das estatísticas descritivas para as três variáveis desse exemplo. Dessa tabela, podemos verificar que, em média, os escores da higiene foram 1,77 (numa escala de 5) no primeiro dia do festival, mas caíram para 0,96 e 0,98 nos dias dois e três, respectivamente.

As outras medidas importantes para o nosso objetivo são a assimetria e a curtose; ambas apresentam um erro padrão associado. O valor da assimetria e da curtose deverão ser zero em uma distribuição normal. Valores de assimetria positivos indicam uma concentração de valores à esquerda enquanto um valor negativo mostra uma concentração de valores à direita. Valores positivos da curtose indicam uma distribuição pontiaguda e valores negativos indicam uma achatada. Quanto mais distante esses valores estiverem de zero, maior a possibilidade de que os dados não sejam normais. Entretanto, os valores reais da assimetria e curtoses não são, eles próprios, informativos. Precisamos padronizar os valores, isto é, transformá-los em escores-*z*. Vimos na Seção 1.5.3 que um escore-*z* é simplesmente um valor da distribuição que apresenta média igual a 0 e desvio padrão igual a 1. O motivo para a conversão dos valores para escores-*z* é que dessa forma eles estarão padronizados. Assim, podemos converter qualquer variável mensurada em qualquer unidade em escores-*z*. Com essa transformação, é possível comparar valores mesmo que tenham sido mensurados em unidades diferentes. Para transformar um valor em um escore-*z*, basta subtrair a média da distribuição e dividir o resultado pelo desvio padrão da distribuição. Assimetria e curtose são convertidas em escores-*z* da mesma forma.

$$Z_{\text{Assimetria}} = \frac{S - 0}{\text{EP}_{\text{Assimetria}}} \qquad Z_{\text{Curtose}} = \frac{K - 0}{\text{EP}_{\text{Curtose}}}$$

Nas equações acima, os valores de S (Assimetria) e K (Curtose) e seus respectivos erros padrões são fornecidos pelo SPSS. Esses escores-*z* podem ser comparados com os valores que esperaríamos serem obtidos apenas por acaso (isto é, valores conhecidos para a distribuição normal mostrados no Apêndice). Assim, um valor absoluto maior do que 1,96 é significativo com $p < 0,05$, e um valor acima de 2,58 é significativo com $p < 0,01$. Valores absolutos acima de 3,29 são significativos com $p < 0,001$. Amostras grandes fornecerão erros padrão pequenos e, desse modo, quando os tamanhos amostrais são grandes, valores significativos serão obtidos mesmo com pequenos desvios da normalidade. Em muitas amostras, valores acima de 1,96 serão suficientes, no entanto, para amostras grandes esse critério deve ser aumentado para 2,58 e com amostras muito grandes, em virtude do problema dos erros padrão pequenos já comentado, nenhum critério deve ser utilizado! Se tivermos uma amostra grande (200 ou mais), é mais importante verificar visualmente a forma da distribuição e procurar os valores da assimetria e da curtose em vez de calcular suas significâncias.

Para os valores da higiene, o escore-*z* da assimetria é $-0,004/0,086 = 0,047$ no primeiro dia, $1,095/0,150 = 7,300$ no segundo dia e $1,033/0,218 = 4,739$ no terceiro dia. Fica

Material com direitos autorais

Dica da Samanta Ferrinho

- Para saber se a distribuição dos escores é normal precisamos verificar os valores da assimetria e da curtose na saída do SPSS.
- Valores positivos da curtose indicam uma distribuição pontiaguda e negativos, uma achatada.
- Quanto mais longe os valores estiverem de zero maior a probabilidade de que os dados não sejam normalmente distribuídos.

Saída 3.1 do SPSS

***Statistics* (Estatísticas)**

		Hygiene (Day 1 of Glastonbury Festival (Higiene do Primeiro dia do Festival de Glastonbury)	*Hygiene (Day 2 of Glastonbury Festival* (Higiene do Segundo dia do Festival de Glastonbury)	*Hygiene (Day 1 of Glastonbury Festival* (Higiene do Terceiro dia do Festival de Glastonbury)
N	*Valid* (Válidos)	810	264	123
	Missing (Desconhecidos)	0	546	687
Mean (Média)		1.7711	0.9609	0.9765
Std. Error of Mean (Erro Padrão da Média)		0.02437	0.04436	0.06404
Median (Mediana)		1.7900	0.7900	0.7600
Mode (Moda)		2.00	0.23	0.44
Standard Deviation (Desvio Padrão)		0.69354	0.72078	0.71028
Variance (Variância)		0.48100	0.51952	0.50449
Skewness (Assimetria)		0.004	1.095	1.033
Std. Error of Skewness (Erro Padrão da Assimetria)		0.086	0.150	0.218
Kurtosis (Curtose)		0.410	0.822	0.732
Std. Error of Kurtosis (Erro Padrão da Curtose)		0.172	0.299	0.433
Range (Amplitude)		3.67	3.44	3.39
Minimum (Mínimo)		0.02	0.00	0.02
Maximum (Máximo)		3.69	3.44	3.41

Existem múltiplas modas. O menor valor é mostrado.

claro então, que embora no primeiro dia os escores sejam simétricos, nos dias 2 e 3 eles apresentam uma assimetria positiva significativa mostrados pelos escores-*z* acima de 1,96 (como já tinha ficado evidente pelos histogramas). No entanto, lembre o que foi dito sobre grandes amostras! Os escores-*z* da curtose são: $-0{,}410/0{,}172 = -2{,}38$ no primeiro dia, $0{,}822/0{,}299 = 2{,}75$ no segundo e $0{,}732/0{,}433 = 1{,}69$ no terceiro dia. Esses valores indicam curtose significativa (ao nível de 5%) em todos os três dias, mas em virtude da amostra grande isso não é uma surpresa e podemos ficar aliviados porque os três valores estão abaixo do nosso valor limite de 3,29.

A saída fornece a distribuição de frequências tabelada de cada variável (não reproduzidas aqui). Essas tabelas listam cada escore e o número de vezes que cada um aparece no conjunto de dados. Além disso, cada valor da frequência é expresso como um percentual do total da amostra. Também é fornecido o percentual acumulado, que nos informa quantos casos (em percentual) estão abaixo de um determinado valor. Assim, por exemplo, podemos ver que somente 15,4% dos escores da higiene

Material com direitos autorais

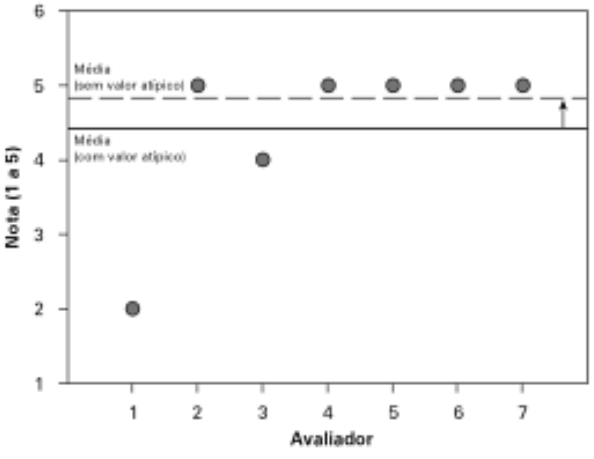
estão abaixo de 1 no primeiro dia do festival. Compare isso com a tabela do segundo dia: 63,3% dos escores foram menores do que 1!

Tendo verificado a forma da distribuição, precisamos procurar valores atípicos (*outliers*) (veja o Quadro 3.1). Existem duas formas de fazer isso: (1) olhar o diagrama de caixa e bigodes e (2) olhar os escores-*z*. Deixarei que o Alex Esperto lide com os escores-*z* no Quadro 3.2 e ficarei com os belos diagramas que são mais adequados ao meu nível de conhecimento. Já mostramos que o SPSS pode criar

Quadro 3.1

Valores atípicos ①

Um valor atípico (*outlier*) é um escore bastante diferente do resto dos dados. Quando analisamos dados devemos estar cientes de tais valores porque eles podem introduzir tendenciosidades no modelo que ajustamos aos dados. Um bom exemplo dessa tendenciosidade pode ser vista em um modelo estatístico simples como a média. Field e Hole (2003) apresentam um exemplo da avaliação da primeira edição deste livro na Amazon.co.uk. Esses valores podem variar de uma a cinco estrelas. Quando ele foi lançado, sete pessoas fizeram uma avaliação e deram uma nota para a primeira edição com valores variando (na ordem em que foram obtidos) de 2, 5, 4, 5, 5, 5, 5. Todas exceto uma dessas avaliações foram semelhantes (principalmente 4 e 5), mas o primeiro valor, de 2, foi bastante diferente dos restantes. No gráfico são mostrados os sete avaliadores no eixo horizontal e as notas atribuídas no eixo vertical. Nesse gráfico existe uma linha horizontal que representa a média dos sete valores (4,43 como verificado). Deve ficar claro que todos os escores, exceto um, ficam próximos dessa linha. O valor 2 é bastante diferente e está situado bem abaixo da média. Esse escore é um exemplo de valor atípico (*outlier*). A linha horizontal pontilhada representa a média dos escores (4,83) quando o valor atípico não é incluído. A linha é mais alta do que a original, indicando que se ignorarmos o escore atípico, a média aumenta (aumenta em 0,4). Isso mostra como um único valor pode viezar a média; nesse caso, o primeiro escore (de dois) pode baixar a média. Em termos práticos, isso tem uma implicação grande porque a Amazon arredonda para meios valores, assim, esse único valor faz diferença entre o valor registrado pela empresa de 4,5 e o valor de 5 estrelas que seria obtido sem a consideração do valor atípico (embora eu esteja ressentido com isso, agradeço o ótimo exemplo que eles estão me fornecendo um ótimo exemplo de um valor atípico!)



(Os dados desse exemplo foram obtidos no *site* http://www.amazon.co.uk/.)

Material com direitos autorais

diagramas de caixa e bigodes na Figura 3.3; repita esse processo para os dados da higiene. O diagrama resultante agora deve parecer diferente (lembre que alteramos o valor incorreto) e ele é mostrado na Figura 3.7.

Os diagramas de caixa e bigodes na Figura 3.7 nos informam sobre a distribuição dos escores da higiene dos três dias do Festival de Glastonbury. O diagrama de caixa e bigodes nos mostra o menor escore (a linha horizontal inferior em cada figura) e o maior (a linha superior horizontal de cada figura). A distância entre a linha mais inferior e a aresta inferior da caixa é a amplitude onde os 25% escores mais baixos podem ser encontrados (é denominado *quartil inferior*). A caixa mostra os 50% dos escores situados no meio do conjunto de valores (denominado ***intervalo interquartílico***): isto é, 50% dos escores são maiores do que a parte mais baixa da figura, mas menores do que a parte superior da figura. A distância entre a aresta superior da caixa e a linha horizontal superior mostra o intervalo onde os 25% maiores escores poderão ser encontrados (*quartil superior*). No meio da caixa há uma linha horizontal levemente mais grossa que as arestas (bordas) da caixa. Essa linha representa a **mediana**, que seria o escore do meio se todos os escores da higiene fossem colocados em ordem. Assim, esses diagramas nos mostram o intervalo dos dados, o intervalo onde estão situados os 50% dos valores do meio do conjunto e o escore mediano. Assim como os histogramas, eles também nos informam se a distribuição é simétrica ou não. No primeiro dia, em que a distribuição é simétrica, os bigodes em cada lado apresentam o

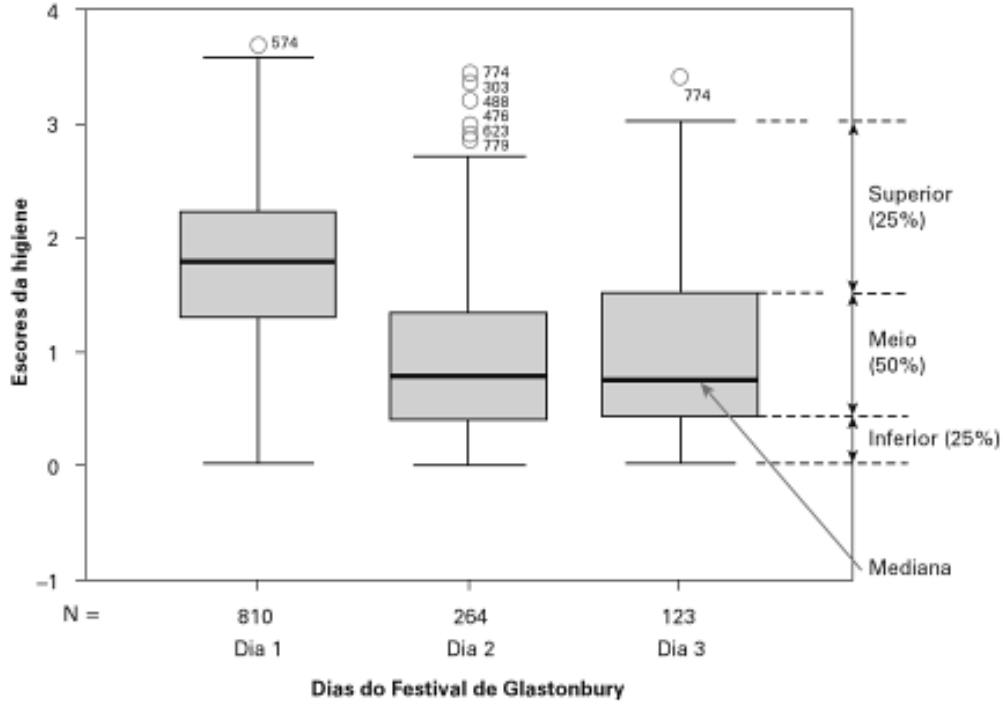


Figura 3.7 Diagrama de caixa e bigodes dos escores de higiene de três dias do Festival de Glastonbury.

Material com direitos autorais

Quadro 3.2 (*Continuação*)

*Isso cria uma nova variável no editor de dados denominada *outlier1*, que contém os valores absolutos dos escores-*z* recém-criados.*

```
RECODE
Outlier1 (3.29 thru Highest = 4) (2.58 thru Highest = 3) (1.96 thru Highest = 2) (Lowest thru 2 = 1).
EXECUTE.
```

Isso recodifica a variável denominada outlier1 de acordo com os pontos de referência que descrevi. Assim, se um valor é maior do que 3,29, ele será codificado como 4, se está entre 2,58 e 3,29, seu código será 3, se estiver entre 1,96 e 2,58, o código é 2, e se for menor do 1,98, será codificado como 1.

```
VALUES LABEL outlier1
1 'Absolute z-score less than 2' 2 'Absolute z-score greater than 1,96' 3 'Absolute z-score greater than 2.58' 4 'Absolute z-score greater than 3.29'.
* Isso atribuirá rótulos apropriados para os códigos definidos acima.*
FREQUENCIES
VARIABLES=outlier1
/ORDER=ANALYSIS.
```

*Finalmente, essa sintaxe utiliza o recurso *frequencies* (frequências) do SPSS para produzir uma tabela nos mostrando o percentual de uns, dois, três e quatros encontrados na variável outlier1.*

Caso atípico 1

		Frequency (Frequências)	*Percent* (Percentual)	*Valid Percent* (Percentual de válidos)	*Percent* (Percentual)
Valid	*Absolute z-score less than 2* (Valor absoluto (Válido) do escore-*z* menor do que 2)	246	30.4	93.2	93.2
	Absolute z-score greater than 1.96 (Valor absoluto do escore-*z* maior do que 1,96)	12	1.5	4.5	97.7
	Absolute z-score greater than 2.58 (Valor absoluto do escore-*z* maior do que 2,58)	4	0.5	1.5	99.2
	Absolute z-score greater than 3.29 (Valor absoluto do escore-*z* maior do que 3,29)	2	0.2	0.8	100.0
	Total	264	32.6	100.0	
Missing	*System* (Sistema desconhecido)	546	67.4		
Total		810	100.0		

A tabela produzida pela sintaxe é mostrada. Olhe para a coluna denominada *"Valid Percent"* (Percentual de válidos). Esperaríamos ver 95% dos casos com valores absolutos menores do que 1,96, 5% (ou menos) com um valor absoluto maior do que 1,96 e 1% (ou menos) com um valor absoluto maior do que 2,58. Finalmente, não esperaríamos casos acima de 3,29 (bem, esses casos são valores atípicos significativos). Para os escores de higiene do segundo dia do Festival de Glastonbury, 93,2% dos valores tem escores-*z* menores do que 1,96. Em outras palavras, 6,8% estão acima (olhando na tabela, obtemos esse valor pela soma de 4,5% + 1,5% + 0,8%). Esse resultado está levemente acima dos 5% esperaríamos fosse normal. Olhando para os valores acima de 2,58, esperaríamos encontrar apenas 1%, mas novamente temos aqui um valor maior, nesse caso, 2,3% (1,5% + 0,8%). Finalmente, descobrimos que 0,8% dos casos estão acima de 3,29 (assim, 0,8% são valores atípicos significativos). Isso sugere que existem muitos valores atípicos nesse conjunto de dados e talvez tenhamos que fazer algo a respeito!

Material com direitos autorais

mesmo tamanho (o intervalo entre os 25% valores superiores e os 25% inferiores são iguais). No entanto, nos dias dois e três, os bigodes situados no topo são maiores do que os situados embaixo, mostrando que a distribuição é assimétrica (isto é, 25% dos escores superiores estão espalhados em um intervalo maior do que os 25% escores inferiores). Finalmente, você terá notado alguns círculos acima de cada um dos diagramas. Esses casos são os valores atípicos (*outliers*). Cada círculo apresenta um número próximo a ele, informando em que linha do editor de dados esse caso poderá ser encontrado. Na próxima seção, veremos o que fazer a respeito desses valores atípicos.

3.3.3 Passo 3: Corrigindo problemas nos dados ②

A última seção nos mostrou várias formas de explorar os nossos dados; vimos como olhar para os problemas com a nossa distribuição de escores e como detectar valores atípicos. A próxima questão é o que fazer a respeito desses problemas. Se você detectar valores atípicos nos dados, existem várias formas de reduzir o impacto desses valores. No entanto, antes que você execute qualquer uma delas, vale à pena verificar se os dados foram digitados corretamente em cada um dos casos problemáticos. Se os dados estiverem corretos, as três principais opções são:

1. **Remover o caso**: Isso irá obrigá-lo a apagar os dados da pessoa que forneceu o valor atípico. Entretanto, isso só deve ser feito se você tiver uma boa razão para acreditar que esse valor não é um representante da população que você pretende amostrar. Por exemplo, se você está investigando fatores que afetam quanto os gatos ronronam e um gato nunca ronrona, isso será um valor atípico (todos os gatos ronronam). Após a inspeção, se você descobrir que esse gato era de fato um cachorro (o motivo dele não ronronar) vestido de gato, teremos motivos para excluir esse caso, porque ele vem de uma população diferente (cachorros que gostam de se vestir como gatos) do que a população-alvo (gatos).
2. **Transformar os dados**: Se você tiver uma distribuição não-normal, isso deverá ser feito de qualquer modo (distribuições assimétricas terão, por natureza, valores atípicos, pois são eles os responsáveis pela falta de simetria!). Tais transformações (ver a seguir) devem reduzir o impacto desses valores.
3. **Substituir o valor:** Se a transformação falhar, você pode considerar a substituição do valor. Colocado dessa forma, isso soa como uma falsificação (você está alterando o valor daquilo que foi efetivamente obtido); no entanto, se o valor que você está alterando não é representativo e ainda está distorcendo o modelo, mudar o valor é o menor dos males! Existem várias opções para substituir um valor:

   (a) O próximo escore mais alto adicionado de um: mude o dado para uma unidade maior do que o próximo valor mais alto do conjunto de dados.

   (b) Inverta o valor do escore-*z*: Um escore-*z* de 3,29 é um valor atípico (veja o Quadro 3.2), assim, podemos calcular qual escore forneceu o valor de 3,29 (ou talvez 3) invertendo a equação do Quadro 3.2 que irá fornecer: $\times = z\ x\ s + \overline{X}$. Tudo isso significa que calculamos a média ($\overline{X}$) e o desvio padrão (s) dos dados, sabemos que z é 3 (precisamente 3,29) e, assim, apenas adicionamos o triplo do desvio padrão à média e substituímos o valor atípico por esse.

   (c) A média mais dois desvios padrão: Uma variação do método acima é utilizar a média mais dois desvios padrão (em vez da média mais três desvios padrão).

Material com direitos autorais

A próxima seção é de arrepiar os cabelos, portanto, não se preocupe se ela não fizer muito sentido. Muitos cursos de graduação não abordam a transformação dos dados assim, sinta-se à vontade para ignorar esta seção se você quiser!

De todas as opções acima, transformar os dados é talvez a melhor, pois em vez de mudar um único valor, uma alteração é feita em todos eles (assim, você não está apenas selecionando um escore para ser alterado, mas fazendo algo para reduzir o impacto dos valores extremos). Dessa forma, a idéia por trás da transformação é mudar todos os dados para corrigir problemas relacionados ao modelo ou valores atípicos. Embora alguns estudantes muitas vezes (compreensivelmente) pensem que transformar os dados seja errado (a frase "falsificar os resultados" vem à mente de algumas pessoas!), de fato, não é isso, pois você faz a mesma coisa com todos os dados.[3] Assim, a transformação dos dados não irá alterar o relacionamento entre variáveis (as diferenças relativas entre pessoas para uma dada variável permanecem as mesmas), mas ela altera as diferenças entre as variáveis (porque altera as unidades de medida). Dessa forma, mesmo se você tiver apenas uma variável que apresenta uma distribuição assimétrica, ainda deve transformar qualquer outra variável do seu conjunto de dados se irá comparar diferenças entre essa variável e outras que pretende transformar. Assim, por exemplo, com os dados da higiene, você pode querer observar como os níveis de higiene se alteraram ao longo dos três dias (isto é, comparar a média do dia 1 com as médias dos dias 2 e 3). Os dados dos dias 2 e 3 são assimétricos e precisam ser transformados, mas em virtude de podermos querer mais tarde comparar os dados aos do primeiro dia, devemos transformar também os dados do primeiro dia (mesmo que eles não sejam assimétricos). Se não transformarmos os dados do primeiro dia, qualquer diferença nos escores da higiene que encontrarmos entre os dias 1, 2 e 3 será por termos transformado uma variável e não as demais.

Quando você tiver dados com assimetria positiva (como ocorre aqui) que precisam ser corrigidos, existem três transformações que podem ser úteis (como veremos, elas podem ser adaptadas para lidar com dados com assimetria negativa também):[4]

1. **Transformação logarítmica ($\log(X_i)$):** Tomar o logaritmo de um conjunto de números esmaga a cauda direita da distribuição. Portanto, é uma boa maneira de reduzir uma assimetria positiva. No entanto, você não pode obter o logaritmo de zero ou de valores negativos, assim, se os seus dados tendem para zero ou produzem números negativos, é necessário adicionar uma constante a todos os valores antes de realizar a transformação. Por exemplo, se você tiver zeros nos dados, faça $\log(X_i + 1)$, ou se tiver valores negativos, adicione qualquer valor que faça o menor dos valores do conjunto se tornar positivo.
2. **Transformação por radiciação ($\sqrt{X_i}$):** Tomar a raiz quadrada de valores grandes tem um efeito maior do que extrair a raiz de pequenos valores. Consequentemente, tomar a raiz quadrada de cada um dos valores fará com que cada valor grande fique mais próximo do centro, parecido com a transformação logarítmica. Isso poderá ser útil para reduzir dados com assimetria positiva; no entanto, você ainda terá o mesmo problema com os números negativos (eles não apresentam raízes quadradas).

[3] Embora não existam conseqüências estatísticas em transformar os dados, como Grayson (2004) aponta, podem existir implicações empíricas ou científicas que se sobreponham os benefícios estatísticos. Especificamente, transformar significa que estamos agora trabalhando em um construto diferente do que foi originalmente mensurado e isso acarreta implicações óbvias na interpretação desses dados.

[4] Você notará, nesta seção, que continuo escrevendo X_i. Vimos no Capítulo 1 que ele se refere ao escore da i-ésima pessoa (assim, o i pode ser substituído pelo nome dessa pessoa em particular, por exemplo, Gregório, $X_i = X_{\text{Gregório}}$ = Escore do Gregório, e para Carol, $X_i = X_{\text{Carol}}$ = Escore da Carol).

Material com direitos autorais

3. **Transformação recíproca ($1/X_i$):** Dividir 1 por cada escore reduz o impacto dos grandes valores. A variável transformada terá um limite inferior de 0 (grandes valores ficarão próximos de zero). Uma coisa deve ser levada em consideração para executar esse tipo de transformação: ela reverte os escores, ou seja, valores que eram originalmente grandes no conjunto de dados se tornarão pequenos no conjunto transformado (próximos a zero), mas escores que eram originalmente pequenos se tornarão grandes após a transformação. Por exemplo, imagine dois escores de 1 e 10. Após a transformação, eles se tornarão $1/1 = 1$ e $1/10 = 0{,}1$: o escore pequeno se torna maior do que o escore grande depois da transformação. No entanto, você pode evitar isso revertendo os escores antes da transformação, encontrando o maior valor e mudando cada escore para o mais alto menos o escore que quer transformar. Assim, a seguinte transformação deverá ser realizada: $1/(X_{\text{Máximo}} - X_i)$.

Cada uma dessas transformações também pode ser utilizada para corrigir dados com assimetria negativa, mas primeiro será necessário reverter os valores (isto é, subtrair cada um dos valores do valor máximo observado). Se você fizer isso, não se esqueça de reverter os valores novamente depois!

3.3.4 Passo 4: Transformando dados utilizando o SPSS ②

O próximo assunto é como realizar essas transformações no SPSS, e isso nos leva ao comando *compute* (calcular). Esse comando permite executar várias funções nas colunas dos dados no editor. Algumas funções típicas são adicionar valores de várias colunas, calcular a raiz quadrada dos valores de um determinada coluna, ou calcular a média de diversas variáveis (colunas). Para obter a caixa de diálogo do comando, utilize o mouse para especificar **Transform⇒Compute…** (Transformar⇒Calcular…). A caixa de diálogo resultante é mostrada na Figura 3.8. Nessa caixa de diálogo, existe uma lista de funções no painel do lado direito e um teclado semelhante a uma calculadora no centro. A maioria das funções da calculadora são óbvias e as mais comuns são:

- **Addition** (Adição)**:** Esse botão coloca um símbolo de mais na área de comandos. Por exemplo, com nossos dados da higiene, "day1 + day2" cria uma coluna onde cada linha contém um escore da higiene da coluna denominada *day1* somado ao escore da coluna denominada *day2* (assim, para o primeiro participante: 2,65 + 1,35 = 4).
- **Subtraction** (Subtração)**:** Esse botão coloca um sinal de menos na área de comandos. Por exemplo, se quisermos calcular a mudança ocorrida na higiene do primeiro dia em relação ao segundo dia, podemos digitar "day2 – day1". Isso criará criar uma coluna onde cada linha contém um valor da coluna denominada *day2* subtraído do valor da coluna denominada *day1* (assim, para o participante 1: 2,65 – 1,35 = −1,30). Dessa forma, a higiene da pessoa caiu 1,30 (na nossa escala de cinco pontos) do primeiro dia para o segundo dia do festival.
- **Multiply** (Multiplicar). Esse botão coloca um sinal de multiplicação na área de comandos. Por exemplo, "day1*day2" criar uma coluna que conterá os escores da coluna denominada *day1* multiplicado pelo escore da coluna denominada *day2* (assim, para o participante 1: 2.65x1,35 = 3,58).
- **Divide** (Dividir). Esse botão coloca um sinal de divisão na área de comandos. Por exemplo, "day1/day2" criar uma coluna que conterá os escores da coluna denominada *day1* divididos pelos escores da coluna denominada *day2* (assim, para o participante 1: 2,65/1,35 = 1,96).
- **Exponentiation** (Exponenciação). Esse botão é utilizado para elevar o termo a uma potência especificada. Assim, "day1**2" cria uma coluna que conterá os escores da coluna denominada *day1*

Material com direitos autorais

elevada à potência 2 (assim, para o participante 1: $2{,}65^2 = 7{,}02$). Da mesma forma "day1**3" cria uma coluna com os valores da variável *day1* ao cubo.

Less than (Menor que). Essa operação é normalmente utilizada pela função *include case* (incluir casos). Se você clicar no botão If..., uma caixa de diálogo aparece e permite que sejam selecionados certos casos nos quais é possível executar alguma operação. Assim, se for digitado "day1 < 1", o SPSS executará a função *compute* (calcular) apenas para aqueles participantes cujo escore de higiene no primeiro dia do festival sejam menores do que 1 (isto é, se **day1** for 0,99 ou menos). Dessa forma, podemos utilizá-la para ver quais pessoas já estavam fedendo no primeiro dia do festival!

Less than or equal to (Menor que ou igual a). Essa operação é a mesma que a anterior, mas os casos que são exatamente 1 também seriam incluídos.

More than (Maior que). Essa operação é normalmente utilizada para incluir casos acima de um certo valor. Assim, se você clicar no botão If... e digitar "day1 > 1", o SPSS fará uma análise apenas nos casos em que o escore da higiene no primeiro dia do festival for maior do que 1 (isto é, 1,01 ou acima). Isso pode ser utilizado, caso você queira, para excluir aquelas pessoas que já estavam fedendo quando o festival começou (como estamos interessados nos efeitos do festival na higiene das pessoas, esses casos iriam contaminar os dados – sem mencionar nossas narinas!)

More than or equal to (Maior que ou igual a). Essa operação é a mesma que a anterior, mas irá incluir casos que são exatamente iguais a 1.

Equal to (Igual a). Você pode utilizar essa operação para incluir casos em que os participantes tenham um valor específico. Assim, "day1 = 1" digitado na caixa de diálogo *if* (se) assegura que somente casos que tenham valores iguais a 1 para a variável **day1** serão incluídos. Isso será mais útil quando tivermos uma variável de grupo e quisermos trabalhar com apenas um dos grupos. Por exemplo, se você tiver a variável **música** para agrupar os gêneros musicais dos participantes do festival (veja a Seção 5.10) como 1 = Pagodeiros, 2 = Metaleiros, 3 = *Punks* e 4 = Ecléticos, e digitar "música = 2", a análise será executada somente com aquelas pessoas que se identificam como "metaleiras" (pessoas que ouvem *heavy metal*).

Not equal to (Não igual a). Essa operação irá incluir todos os casos exceto aqueles com valores específicos. Assim, "música ≅ 2" incluirá todos os casos exceto aqueles que se identificaram como metaleiros.

Type&Label... **Variable type** (Tipo de variável). Esse botão abre uma caixa de diálogo que permite fornecer um nome para uma nova variável e especificar que tipo de variável ela é (por exemplo, numérica).

O resultado de qualquer uma das funções de cálculo irá produzir uma nova coluna no editor de dados, portanto, a primeira coisa a fazer é digitar um nome para essa nova variável (na caixa denominada *Target Variable* [variável alvo]). Se você digitar um nome que já existe, o SPSS irá informar e perguntar se quer alterá-lo ou substituir a variável existente. Se você responder "Yes" (Sim), o SPSS substituirá os dados existentes na coluna com os resultados da função utilizada. Se você responder "No" (Não), nada acontecerá e você precisará renomear a nova variável. Se você clicar em Type&Label..., outra caixa de diálogo aparecerá, em que você pode fornecer à nova variável um rótulo descritivo e especificar se ela é numérica ou texto (veja a Seção 2.4.5). O quadro denominado *Numeric Expression* (Expressão Numérica) é o espaço onde os comandos numéricos são digitados (chamei esse espaço de "área de comandos"). Você poderá entrar com

Material com direitos autorais

mos adicionar uma constante à nossa variável da mesma forma que foi feito com a transformação logarítmica. Qualquer constante pode ser utilizada, mas 1 é um valor conveniente para esses dados. Assim, em vez de selecionar a variável que queremos transformar, clique em [(]. Isso acrescentará parênteses ao quadro denominado *numeric expression* (expressão numérica), assim, assegure-se de que o cursor esteja entre os parênteses e selecione a variável que você quer transformar da lista e transfira para o painel utilizando a seta [▸] (ou digite o nome da variável manualmente). Agora, clique em [+] e então em [1] (ou digite + 1 utilizando o teclado). O quadro denominado *Numeric Expression* (expressão numérica) deve agora conter o seguinte texto *1/(day1 + 1)*. Clique em [OK] para criar a nova variável contendo os valores transformados. Repita esse processo para os dados dos dois dias restantes. Se você estiver calculando muitas variáveis, poderá ser mais rápido utilizar a janela de sintaxe como explicado no Quadro 3.3.

Vamos dar uma olhada nas nossas novas distribuições (Figura 3.9). Compare-as com as distribuições originais (não transformadas) na Figura 3.5. Agora, você pode ver que todas as três transformações limparam os dados dos escores da higiene para o dia 2: a assimetria positiva desapareceu (a transformação pela raiz

Quadro 3.3

Utilizando sintaxe para calcular novas variáveis ③

Se você estiver calculando muitas variáveis, pode ser mais rápido utilizar a sintaxe – por exemplo, para criar os dados transformados no exemplo deste capítulo. No *site* www.artmed.com.br, há um arquivo de sintaxe, **Transformations.sps**, que escrevi para todas as nove transformações discutidas. Carregue esse arquivo e rode a sintaxe ou abra a janela de sintaxe (veja a Seção 2.6) e digite o seguinte:

```
COMPUTE logday1 = LG10(day1 + 1).
COMPUTE logday2 = LG10(day2 + 1).
COMPUTE logday3 = LG10(day3 + 1).
COMPUTE sqrtday1 = SQRT(day1).
COMPUTE sqrtday2 = SQRT(day2).
COMPUTE sqrtday3 = SQRT(day3).
COMPUTE recday1 = 1/(day1 + 1).
COMPUTE recday2 = 1/(day2 + 1).
COMPUTE recday3 = 1/(day3 + 1).
EXECUTE
```

Cada comando COMPUTE (Calcular) faz o mesmo que você faria utilizando a caixa de diálogo na Figura 3.8. Assim, as três primeiras linhas simplesmente solicitam que o SPSS crie três novas variáveis (**logay1**, **logday2** e **logday3**), que são as transformações logarítmicas das variáveis day1 (dia 1), day2 (dia 2) e day3 (dia 3) mais 1. As próximas três linhas fazem a mesma coisa só que agora utilizando a função SQRT e, assim, calculam a raiz quadrada de **day1**, **day2** e **day3** criando as novas variáveis **sqrtday1**, **sqrtday2** e **sqrtday3**, respectivamente. As próximas três linhas executam a transformação recíproca da mesma forma que as outras 2 funções fizeram. A linha final apresenta o comando EXECUTE (Executar) sem o qual nenhum dos anteriores será processado! Note ainda que cada uma das linhas termina com um ponto final.

Material com direitos autorais

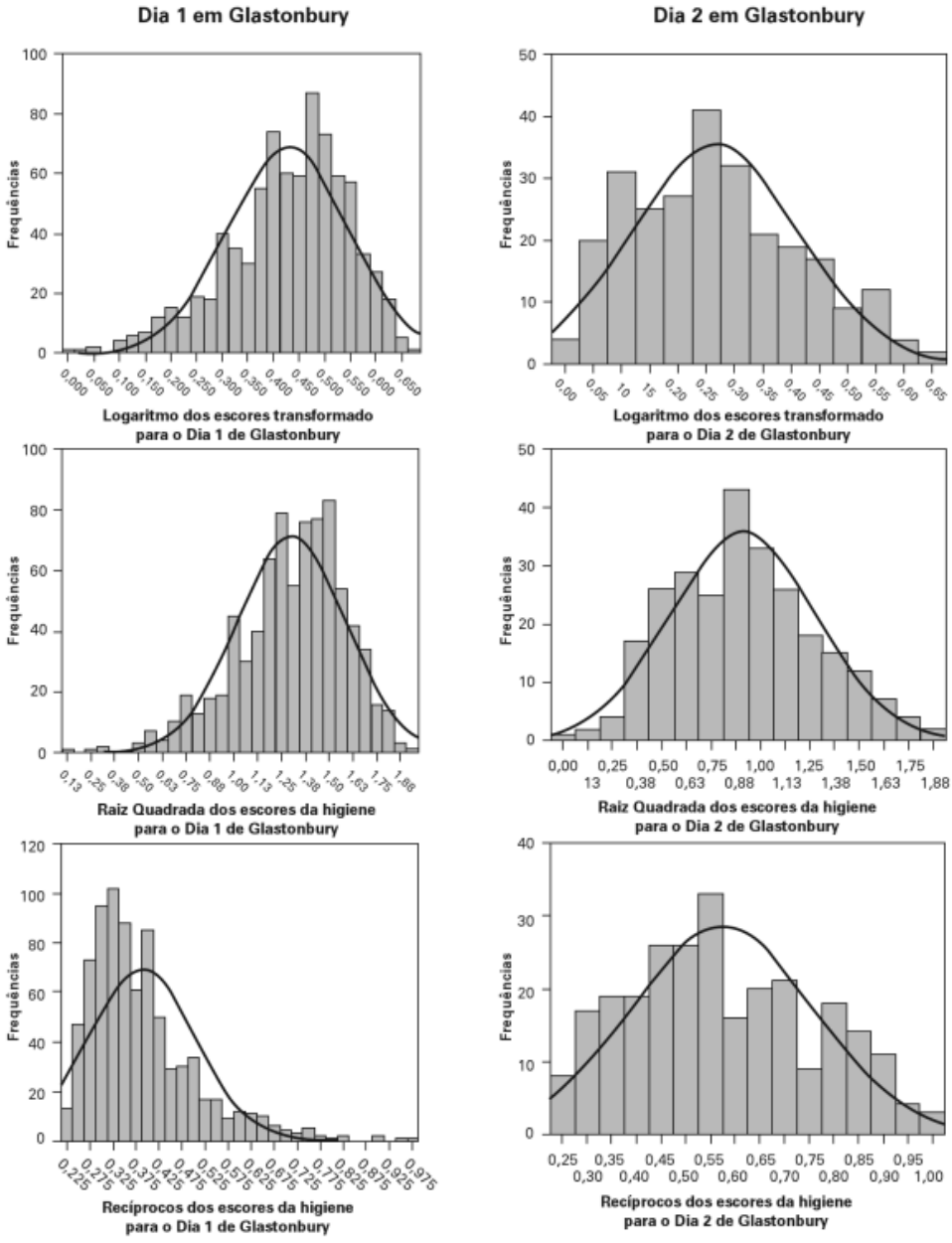


Figura 3.9 Distribuição dos dados da higiene do dia 1 e do dia 2 após as transformações.

quadrada, em particular, foi útil). No entanto, em virtude dos nossos escores da higiene no dia 1 não terem assimetria significativa, eles agora apresentam uma leve assimetria negativa para a transformação logarítmica e da raiz quadrada e assimetria positiva para a transfor-

Material com direitos autorais

mação recíproca![5] Se estivermos utilizando apenas os escores do segundo dia, podemos utilizar os dados transformados; no entanto se quisermos verificar os dois dias em conjunto, teremos que ponderar se os benefícios da transformação sobre os dados do segundo dia não são sobrepujados pelos problemas criados com os dados do primeiro dia – a análise de dados pode ser frustrante às vezes!

Se a transformação corrige os problemas, execute qualquer análise estatística que você precise sobre os escores transformados. Se ela não corrigir o problema, você pode considerar a utilização de um teste que não tenha por base a hipótese de dados normalmente distribuídos e, à medida que você percorrer os vários capítulos do livro apresentarei esses testes.[6]

3.4 EXPLORANDO GRUPOS DE DADOS ①

Algumas vezes temos dados em que existem diferentes grupos de pessoas (homens e mulheres, universidades diferentes, pessoas com e sem depressão, por exemplo). Existem várias maneiras de produzir estatísticas básicas para grupos separados de pessoas (e veremos esses métodos na próxima seção). No entanto, pretendo utilizar essa oportunidade para lhe apresentar uma função denominada *split file* (dividir arquivo), que permite repetir qualquer análise sobre vários grupos de casos. A função *split file* (dividir arquivo) permite especificar uma variável de grupo (lembre que essas variáveis são utilizadas para especificar categorias de pessoas). Qualquer procedimento posterior no SPSS irá ser executado em cada uma das categorias especificadas pela variável de grupo.

Você provavelmente está cansado dos dados sobre a higiene do festival de Glastonbury, portanto, vamos utilizar o arquivo de dados **SPSSExam.sav**. Esse arquivo contém dados do desempenho de estudantes em um exame sobre o SPSS. Quatro variáveis foram mensuradas: **exam (exame)** (escores dos exames sobre o SPSS de alunos do primeiro ano em percentagem), **computer (computador)** (o conhecimento computacional mensurado em percentual), **lecture (aulas)** (percentual de aulas sobre o SPSS frequentadas) e **numeracy (numerácia)** (uma medida da habilidade numérica em uma escala de 1 a 15). Existe ainda uma variável denominada **uni** indicando se o estudante frequenta a Universidade de Sussex ou a de Duncetown*. Para iniciar, abra o arquivo **SPSSExam.sav** (veja a Seção 2.8 se não lembrar como abrir um arquivo). Vamos começar olhando os dados como um todo.

3.4.1 Executando a análise para todos os dados ①

Para olhar a distribuição das variáveis, podemos utilizar o comando *frequencies* (frequências), apresentado na seção anterior (veja a Figura 3.6). Utilize essa caixa de diálogo e coloque todas as quatro variáveis (**exam, computer, lecture** e **numeracy**) no quadro *variable (s)* (variável [is]). Clique em Statistics... para selecionar a caixa de diálogo *statistics* (estatísticas) e selecione algumas medidas de tendência central (média, moda e mediana), medidas de variabilidade (amplitude, desvio padrão, variância e quartis) e medidas de forma (assimetria e curtose). Clique também em Charts para acessar a caixa de diálogo dos diagramas e selecione a distribuição de frequências com a curva normal (veja a Figura 3.6 se precisar

[5] A inversão da assimetria que aconteceu com a transformação recíproca ocorre, como mencionada antes, porque essa transformação inverte o valor dos escores.

[6] Por conveniência, muitos livros-texto se referem a esses testes como *não-paramétricos* ou *distribuições livres* e os colocam em um capítulo separado. De fato, nenhum desses termos é particularmente preciso (pois nenhum desses testes está livre de suposições), mas, para manter a tradição, eu os coloquei em um capítulo (13) próprio, no ostracismo e solitários.

* N. de T.: É um nome fictício, atribuído pelo autor, que em português significa cidade dos burros.

Material com direitos autorais

de ajuda com essa opção). Retorne à caixa de diálogo principal e clique em Continue e, uma vez nela, clique em OK para executar a análise.

3.4.2 Saída do SPSS para todos os dados ①

A saída 3.2 do SPSS mostra a tabela das estatísticas descritivas para as quatro variáveis deste exemplo. A partir dessa tabela, é possível ver que, em média, os estudantes frequentaram aproximadamente 60% das aulas, eles obtiveram 58% no exame do SPSS, tiveram somente 51% no teste de conhecimento ou habilidade computacional e apenas 5 em uma escala até 15 em numerácia, isto é, habilidade numérica. Além disso, o desvio padrão para a habilidade computacional foi relativamente pequeno comparado ao do percentual de aulas freqüentadas e as notas do exame. Essas últimas duas variáveis apresentaram várias modas (multimodais). As outras medidas importantes são a assimetria e a curtose, ambas apresentando um erro padrão associado. Já discutimos essas medidas anteriormente e verificamos que podemos convertê-las em escores-z dividindo-as pelos seus erros padrões. Para os escores do exame, o escore-z da assimetria é $-0{,}107/0{,}241 = -0{,}44$. Para a numerácia, o escore-z da assimetria foi de $0{,}961/0{,}241 = 3{,}99$. Fica evidente que os escores da numerácia apresentam assimetria positiva, indicando uma concentração de valores na parte esquerda da distribuição (isto é, muitos estudantes pontuaram mal). Tente calcular o escore-z para as demais variáveis.

A saída fornece tabelas das distribuições de freqüências de cada uma das variáveis (que não são apresentadas aqui). Essas tabelas listam cada escore e o número de vezes que ele ocorreu no conjunto de dados. Adicionalmente, cada valor da freqüência é expresso como um percentual da amostra (nesse caso, as frequências e as percentagens são as mesmas, pois o tamanho da amostra é igual a 100). Também, a percentagem acumulada é dada, que nos informa quantos casos estão abaixo

Saída 3.2 do SPSS

Statistics **(Estatísticas)**

		Computer literacy (Conhecimento Computacional)	*Percentage on SPSS Exam* (Percentual no Exame do SPSS)	*Percentage of lectures attended* (Percentual de comparecimento às aulas)	*Numeracy* (Numerácia)
N	*Valid* (Válidos)	100	100	100	100
	Missing (Desconhecidos)	0	0	0	0
Mean (Média)		50.7100	58.1000	59.7650	4.8500
Std. Error of Mean (Erro Padrão da Média)		0.8260	2.1316	2.1685	0.2706
Median (Mediana)		51.5000	60.0000	62.0000	4.0000
Mode (Moda)		54.00	72.00[a]	48.50[a]	4.00
Standard Deviation (Desvio Padrão)		8.2600	21.3156	21.6848	2.7057
Variance (Variância)		68.2282	454.3535	470.2296	7.3207
Skewness (Assimetria)		−0.174	−0.107	−0.422	0.961
Std. Error of Skewness (Erro Padrão da Assimetria)		0.241	0.241	0.241	0.241
Kurtosis (Curtose)		0.364	−1.105	−0.179	0.946
Std. Error of Kurtosis (Erro Padrão da Curtose)		0.478	0.478	0.478	0.478
Range (Amplitude)		46.00	84.00	92.00	13.00
Minimum (Mínimo)		27.00	15.00	8.00	1.00
Maximum (Máximo)		73.00	99.00	100.00	14.00

a. Existem múltiplas modas. O menor valor é mostrado.

Material com direitos autorais

de certo valor. Assim, por exemplo, podemos ver que 66% dos escores da numerácia foram 5 ou menos, 74% foram 6 ou menos, e assim por diante. Olhando em outra direção, podemos perceber que somente 8% (100% − 92%) tiveram um escore maior do que 8.

Finalmente, foram obtidos os histogramas de cada uma das variáveis sobrepostos por uma distribuição normal. Esses gráficos são apresentados na Figura 3.10 e nos mostram várias coisas. Primeiro, parece que o conhecimento computacional tem distribuição aproximadamente normal (poucas pessoas são muito boas e poucos são muitos ruins com computadores, a maioria tem um grau de conhecimento similar). Os escores do exame são bem interessantes porque essa distribuição é claramente não-normal; de fato, ela parece suspeitamente bimodal (existem dois picos indicando duas modas). Essa observação corresponde a uma anterior da tabela das estatísticas descritivas. O comparecimento às aulas é aparentemente normal, mas as caudas da distribuição são bastante pesadas (isto é, embora muitas pessoas compareçam à maioria das aulas, existe um número de almas dedicadas que comparecem a todas e um grupo maior do que o "normal" que comparecem a poucas). Esse é o motivo dela apresentar altas freqüências nas duas caudas. Finalmente, o teste de numerácia produziu dados com uma assimetria positiva acentuada (isto é, a maioria das pessoas foi muito mal nesse teste e poucos foram bem). Isso corresponde ao que a estatística de assimetria já havia indicado.

Estatísticas descritivas e histogramas são boas formas de obter um retrato instantâneo

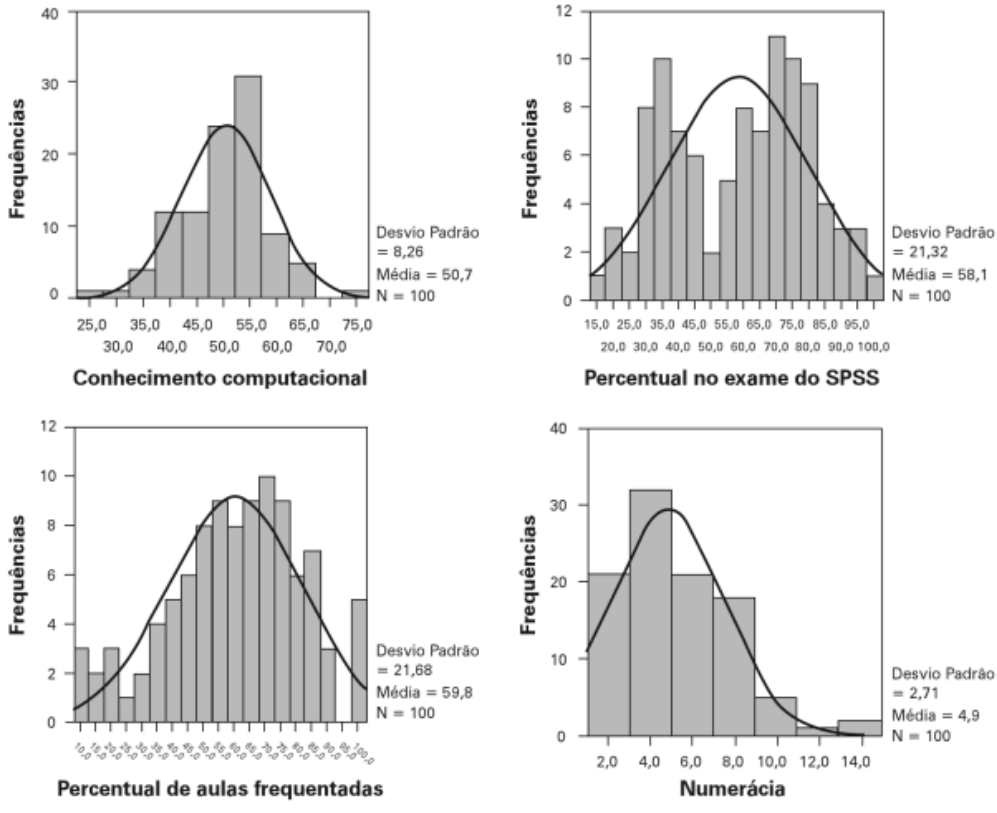


Figura 3.10 Histograma dos dados do exame sobre o SPSS.

Material com direitos autorais

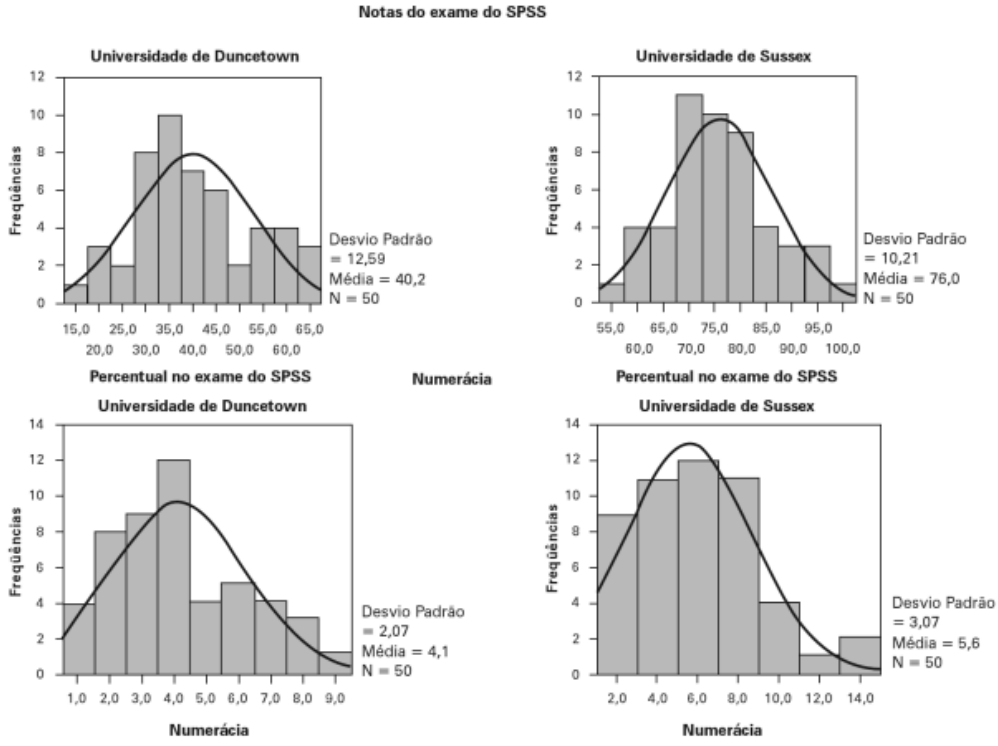


Figura 3.12 Distribuições dos escores no exame e na numerácia para os estudantes das universidades de Duncetown e Sussex.

3.5 TESTANDO SE UMA DISTRIBUIÇÃO É NORMAL ①

Olhar para os histogramas é uma opção, mas eles nada informam sobre a distribuição estar próxima o suficiente da normalidade. A análise do histograma é subjetiva e aberta a abusos (Posso imaginar pesquisadores sentados olhando para uma distribuição completamente distorcida e dizendo "bem Carlos, ela parece normal para mim" e Carlos replicando "com certeza"). O que é necessário é um teste objetivo para decidir se uma distribuição é ou não-normal. As estatísticas de assimetria e curtose que vimos antes informam um pouco sobre desvios da normalidade, mas cada uma delas lida com apenas um aspecto da não-normalidade. Outra maneira de tratar o problema é ver se a distribuição como um todo se desvia de uma distribuição normal modelo. Os testes de Kolmogorov-Smirnov (Figura 3.13) e de Shapiro-Wilk fazem justamente isso: eles comparam escores de uma amostra a uma distribuição normal modelo de mesma média e variância dos valores encontrados na amostra. Se o teste é não-significativo ($p > 0,05$), ele nos informa que os dados da amostra não diferem significativamente de uma distribuição normal (isto é, eles podem ser normais). Por outro lado, se o teste é significativo ($p < 0,05$), a distribuição em questão é significativamente diferente de uma distribuição normal (isto é, ela é não-normal). Esses testes parecem ótimos: em um procedimento eles nos dizem se os nossos escores são normalmente distribuídos (beleza!).

Material com direitos autorais

Figura 3.13 Andrei Kolmogorov, querendo ter uma Smirnov.

Entretanto, eles apresentam limitações porque com amostras grandes é muito fácil obter valores significativos a partir de pequenos desvios da normalidade e, assim, um resultado significativo não necessariamente nos informa se o desvio da normalidade é suficiente para prejudicar os procedimentos estatísticos que serão aplicados aos dados. Acredito que o sensato a fazer é: de qualquer forma utilize esses testes, mas faça diagramas dos dados bem como tente obter uma decisão sobre a extensão da não-normalidade.

3.5.1 Executando o teste de Kolmogorov-Smirnov no SPSS ①

O teste de Kolmogorov-Smirnov (K-S de agora em diante) pode ser acessado pelo comando *explore* (explorar) **Analyze⇒Descriptive Statistics⇒Explore...** (Analisar⇒Estatística Descritiva⇒Explorar...),[7] A Figura 3.14 mostra a caixa de diálogo para o comando *explore* (explorar). Primeiro, entre com qualquer variável de interesse no quadro denominado *Dependent list* (Lista Dependente) selecionando-as no lado esquerdo e transferindo para o direito clicando em ▶. Para esse exemplo, apenas selecione os valores das variáveis exame e numerácia. É também possível selecionar um fator (variável de grupo) pelo qual podemos dividir a saída – (assim, se você selecionar **uni** e transferi-la para o quadro denominado *Factor List* [Lista de fatores], o SPSS irá produzir uma análise exploratória para cada um dos grupos – parecido com o comando *split file* [dividir arquivo]. Se você clicar em Statistics..., uma caixa de diálogo aparece, mas a opção por omissão é suficiente (produzirá médias, desvios padrão e assim por diante). A opção mais interessante para o nosso objetivo é acessada clicando em Plots..., que abrirá uma nova caixa de diálogo onde você pode selecionar a opção P Normality plots with tests; isso irá produzir tanto um teste K-S quanto um gráfico denominado *normal Q-Q plots* (diagramas Q-Q normais)* para todas as variáveis selecionadas. Por omissão, o SPSS também irá apresentar diagramas de caixa e bigodes. Separado por grupos se um fator foi especificado) e também um diagrama de caule e folhas. Clique em Continue para retornar à caixa de diálogo principal e clique em OK para executar a análise.

3.5.2 Saída do procedimento *Explore* (Explorar) ①

A primeira tabela produzida pelo SPSS contém estatísticas descritivas (médias, etc.) e deve ter os mesmos valores que a tabela obtida com a utilização do procedimento frequências (*frequencies*). A tabela importante é a do teste K-S.

Essa tabela inclui o próprio teste, os graus de liberdade (que devem ser iguais ao tamanho da amostra) e os valores da significância do teste. Lembre que um valor significativo (Sig. < 0,05) indica um desvio da normalidade. Tanto para os valores do exame quanto para os da numerácia o teste K-S é altamente

[7] Esse caminho seria **Statistics⇒Summarize⇒Explore...** (Estatísticas⇒Resumir⇒Explorar...) na versão 8.0 e anteriores.

* N. de T.: Um diagrama Q-Q é um gráfico onde o quantil do primeiro conjunto de dados é apresentado contra o quantil de um segundo conjunto de dados. Por quantil entendemos um determinado percentual de valores abaixo de um dado valor. Alguns quantis especiais são: quartis, decis e percentis.

Material com direitos autorais

***Tests of Normality* (Testes de normalidade)**

	Kolmogorov-Smirnov[a]			*Shapiro-Wilk*		
	Statistic (Estatística)	*Df* (gl)	*Sig.* (Significância)	*Statistic* (Estatística)	*Df* (gl)	*Sig.* (Significância)
Percentage on SPSS exam (Percentual no exame do SPSS)	0.102	100	0.012	0.961	100	0.005
Numeracy (Numerácia)	0.153	100	0.000	0.924	100	0.000

a. Correção de significância de Lilliefors.

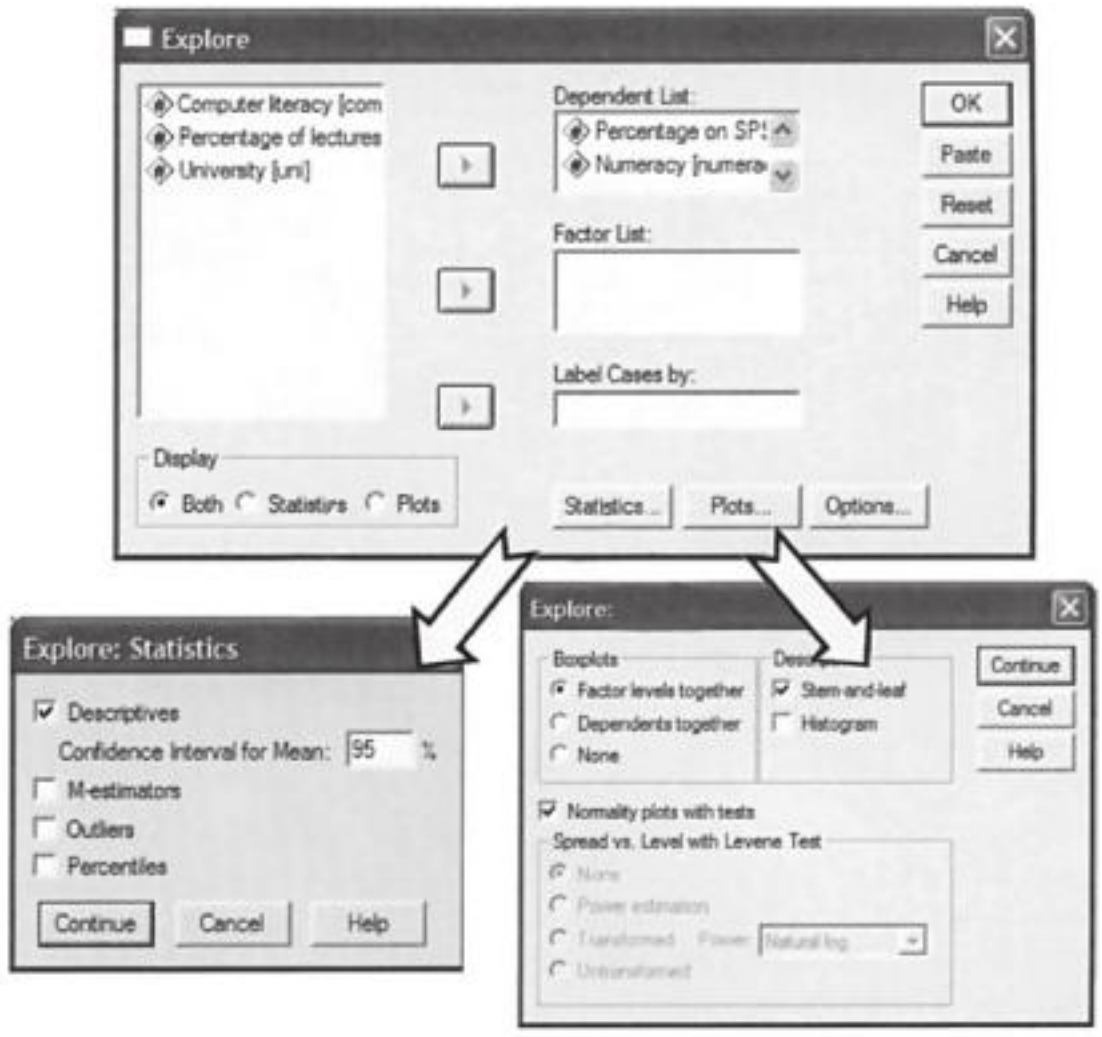


Figura 3.14 Caixas de diálogo para o comando *explore* (explorar).

significativo, indicando que as duas distribuições se desviam da normalidade. Esse resultado provavelmente reflete a distribuição bimodal encontrada para os escores do exame e a distribuição com assimetria positiva observada para os escores da numerácia. Entretanto, esses testes confirmam que esses desvios eram significativos. Essa confirmação é importante porque os histogramas nos informam apenas que nossas distribuições amostrais se desviam da normalidade; eles não nos dizem se esse desvio é importante. A estatística teste para o teste K-S é representada por D, assim, podemos relatar esses resultados da seguinte forma. A percentagem no exame do SPSS, $D(100) = 0{,}10$, $p < 0{,}05$, e para os escores da numerácia, $D(100) = 0{,}15$, $p < 0{,}001$, foram ambos significativamente não-normais. Os

Material com direitos autorais

números nos parênteses são os graus de liberdade (gl) da tabela.

Concluindo, tenha em mente que quando olhamos os escores dos exames considerando grupos separadamente, as distribuições parecem praticamente normais; mas se executarmos testes separados para as duas universidades (colocando **uni** na caixa denominada *Factor List* [Lista de Fatores], como na Figura 3.17) o teste K-S poderá ser não-significativo. De fato, se você fizer isso, obterá a saída 3.4 do SPSS na forma de tabela, que mostra que os percentuais do exame sobre o SPSS são, de fato, normais entre os dois grupos* (os valores na coluna denominada *Sig.* são maiores do que 0,05). Isso é importante porque se nossa análise envolve a comparação entre dois grupos, o que importa não é a distribuição global, mas a distribuição de cada grupo.

O SPSS também produz um diagrama Q-Q normal para qualquer variável especificada (veja a Figura 3.15). O diagrama Q-Q normal traça os valores que esperaríamos obter se a distribuição fosse normal (valores esperados) contra os valores realmente vistos nos dados (valores observados). Os valores esperados são uma linha reta na diagonal do primeiro quadrante do sistema de coordenadas, onde os valores observados (os pontos no diagrama) devem cair exatamente ou bem próximos da linha (significando que os valores observados são os mesmos que se esperaria obter de um conjunto normalmente distribuído). Qualquer

* N. de T.: Como a hipótese nula não está sendo rejeitada aqui, eu colocaria que "as distribuições são possivelmente normais" em vez de "de fato normais", como o autor coloca.

Saída 3.4 do SPSS

***Tests of Normality* (Testes de normalidade)**

		Kolmogorov-Smirnov[a]			*Shapiro-Wilk*		
University (Universidade)		*Statistic* (Estatística)	*Df* (gl)	*Sig.* (Signf.)	*Statistic* (Estatística)	*Df* (gl)	*Sig.* (Signf.)
Percentage on SPSS exam (Percentual no exame do SPSS)	*Duncetown*	0.106	50	0.200*	0.972	50	0.238
	Sussex	0.073	50	0.200*	0.984	50	0.715
Numeracy (Numerácia)	*Duncetown*	0.183	50	0.000	0.941	50	0.015
		0.155	50	0.004	0.932	50	0.007

* Esse é um limite inferior da verdadeira significância.
a. Correção de significância de Lilliefors.

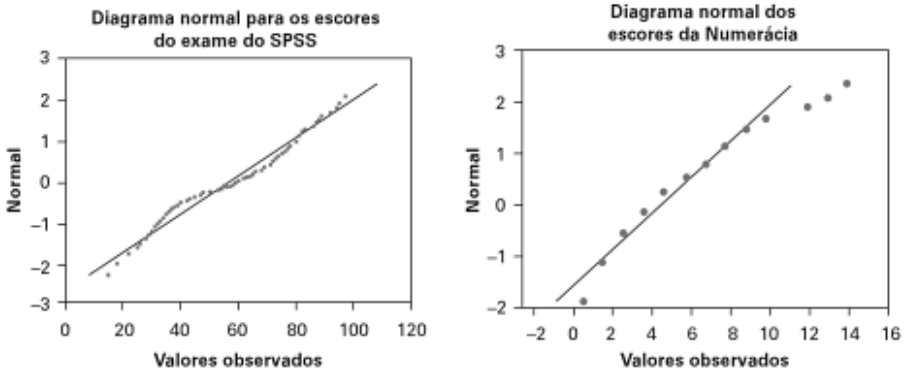


Figura 3.15 Diagramas Q-Q normais das variáveis escores no Exame do SPSS e na numerácia.

Material com direitos autorais

desvio dos pontos em relação à linha reta diagonal representa um desvio da normalidade. Assim, se o diagrama Q-Q se parece com uma linha reta com uma longa cobra serpenteando ao redor, você tem alguns desvios da normalidade! Especificamente, quando a linha fica de forma consistente abaixo da diagonal ou acima, isso está mostrando que a curtose é diferente de uma distribuição normal e, quando os pontos apresentarem uma forma de S, o problema será com a assimetria.

Em ambas as variáveis analisadas já sabemos que os dados são não-normais e esses diagramas confirmam essa observação porque os pontos se desviam substancialmente da linha. Vale à pena notar que os maiores desvios ocorrem com a variável numerácia e isso está de acordo com o valor altamente significativo da variável no teste K-S. Tal desvio da normalidade nos informa que não podemos utilizar um teste paramétrico porque a hipótese de normalidade não se verifica. Nessas circunstâncias, podemos em alguns casos considerar a utilização de um teste não-paramétrico como forma de verificar a hipótese de interesse (veja o Capítulo 4).

3.6 TESTANDO A HOMOGENEIDADE DA VARIÂNCIA ①

Até agora, me concentrei na hipótese de dados normalmente distribuídos; entretanto, no início do Capítulo 1 mencionei outra suposição, a homogeneidade da variância. Essa hipótese significa que à medida que você avança entre os níveis de uma variável, a variância da outra não deve mudar. Se você coletou grupos de dados, isso significa que a variância da sua variável ou variável de saída deve ser a mesma em cada um desses grupos. Se você coletou dados contínuos (como os delineamentos intercorrelacionais), essa suposição significa que a variância de uma variável deve ser estável em todos os níveis da outra variável. Vamos ilustrar com um exemplo. Uma otologista está interessada no efeito dos altos sons dos concertos sobre a audição dos frequentadores. Assim, ela decidiu enviar 10 pessoas a uma turnê com a banda mais barulhenta que ela pode encontrar, Motörhead. Essas pessoas foram a *shows* em Brixton (Londres), Brighton, Bristol, Edinburgo, Newcastle, Cardiff e Dublin. Após cada *show*, a otologista media o número de horas que as pessoas ficavam com os ouvidos zumbindo.

A Figura 3.16 mostra o número de horas que cada pessoa teve zumbido no ouvido após cada concerto (cada pessoa está representada por um círculo). As linhas horizontais representam o número médio de horas que existiu zumbido nos ouvidos após cada *show* e essas médias estão conectadas por uma linha de forma que podemos ver a tendência geral dos dados. Lembre-se de que para cada *show*, os círculos são os escores a partir do qual as médias são calculadas. Agora, podemos ver nos dois gráficos que a média aumenta à medida que as pessoas vão a mais *shows*. Assim, após o primeiro *show* os ouvidos zumbem por aproximadamente 12 horas, mas depois do segundo, eles zumbem por aproximadamente 15 a 20 horas e na noite final da turnê, eles zumbem por 45 a 50 horas (dois dias). Assim, existe um efeito acumulado dos *shows* sobre o zumbido nos ouvidos. Esse padrão é visto nos dois gráficos; a diferença entre os gráficos não está nas médias (que são basicamente as mesmas), mas em termos do espalhamento dos escores em torno da média. Se olharmos para o gráfico da esquerda, veremos que a dispersão dos escores permanece praticamente a mesma após cada *show* (os escores estão na maioria agrupados em torno da média). Colocando de outra forma, se for mensurada a distância vertical entre os escores mais alto e o mais baixo após o *show* de Brixton e fizermos o mesmo após os demais, todas essas distâncias serão aproximadamente iguais. Embora a média aumente, o espalhamento dos escores para a perda de audição é a mesma a cada nível da variável *show* (a dispersão dos escores é a mesma após Brixton, Brighton, Bristol, Edinburgo, Newcastle, Cardiff e Dublin). Isso é o queremos dizer por *homogeneidade da variância*. O gráfico da direita mostra uma situação diferente: se você examinar a dispersão dos escores após o *show*

Material com direitos autorais

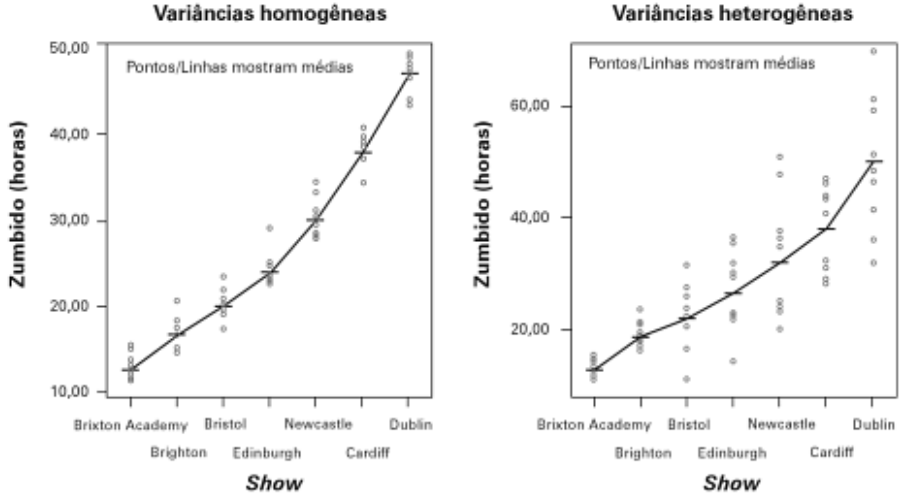


Figura 3.16 Gráficos ilustrando dados com variâncias homogêneas e heterogêneas.

de Brixton, verá que estão bem concentrados em torno da média (a distância vertical entre o maior e o menor escore é pequena), mas após o *show* de Dublin (por exemplo) os escores estão bastante espalhados em torno da média (a distância vertical entre o maior e o menor escore é grande). Esse é um exemplo de *heterogeneidade da variância*; isto é, em alguns níveis da variável *show* a variância dos escores é diferente dos outros níveis (nesse caso, a distância vertical entre o maior e o menor escore é diferente entre diferentes *shows*).

Espero que você tenha captado a idéia do que a homogeneidade da variância significa. Mas como testá-la? Muitos procedimentos estatísticos apresentam particularidades únicas para fazê-lo: na análise correlacional, como a regressão, a tendência é utilizar gráficos (veja a Seção 5.8.7) e para grupos de dados, a tendência é utilizar o **teste de Levene**. O teste de Levene verifica a hipótese de que a variância nos grupos é a mesma (isto é, a diferença entre as variâncias é zero). Dessa forma, se o teste de Levene é significativo quando $p \leq 0{,}05$, podemos concluir que a hipótese nula é incorreta e que as variâncias são significativamente diferentes – assim, a suposição de variâncias homogêneas foi violada. Se, no entanto, o teste de Levene não é significativo (isto é, $p > 0{,}05$), aceitamos a hipótese nula de que as diferenças entre as variâncias é zero – as variâncias são aproximadamente iguais e a suposição é mantida. Embora o teste de Levene possa ser selecionado como uma opção entre muitos testes estatísticos, ele também pode ser uma alternativa quando você está explorando os dados (e é melhor fazê-lo agora do que esperar até a análise principal).

Como ocorreu com o teste K-S (e outros de normalidade), quando o tamanho da amostra é grande, pequenas diferenças entre as variâncias dos grupos podem produzir um teste de Levene significativo (em virtude do que vimos no Capítulo 1, o poder do teste é aumentado). Uma dupla checagem útil é olhar a **razão das variâncias**. Isso é a razão das variâncias entre os grupos de pessoas com a maior variância e a do grupo com a menor variância. Assim, se você olhar para as variâncias entre grupos diferentes, pegue o valor mais alto que encontrar e divida pelo valor mais baixo que encontrar. Se o resultado (razão) for menor que 2, é seguro afirmar que há homogeneidade das variâncias.

É possível realizar o teste de Levene utilizando o menu *explore* (explorar) que foi usado na seção anterior. Para esse exemplo, pegaremos os dados do exame do SPSS que foram

Material com direitos autorais

como utilizar alguns dos procedimentos gráficos do SPSS para criar gráficos de médias (isso é útil quando você fizer o relatório ou quando for apresentá-lo a outras pessoas). Como você irá criar esses gráficos no SPSS depende em grande parte da forma que os dados foram coletados (se foram utilizadas as mesmas pessoas ou pessoas diferentes para medir os diferentes níveis de cada variável). Para começar, imagine que o diretor de uma companhia cinematográfica está interessado em verificar se existe algo conhecido como "filme de garotas" (um filme que tipicamente interessa mais às mulheres do que aos homens). Ele pegou 20 homens e 20 mulheres e mostrou à metade de cada amostra um filme supostamente "de garotas" (*O diário de Bridget Jones*) e para a outra metade um filme que não se encaixaria nessa categoria (*Amnésia* – um filme brilhante, por sinal). Em todos os casos ele mensurou a excitação fisiológica como uma medida de quanto o filme foi apreciado. Os dados estão no arquivo denominado **ChickFlick.sav** no *site* www.artmed.com.br. Carregue esse arquivo agora.

Para traçar um diagrama das médias, iremos utilizar os gráficos interativos do SPSS. A principal vantagem desses gráficos é que é possível colocar barras de erros (veja o Capítulo 8) e eles podem ser atualizados mesmo quando a base de dados estiver fechada (veja o Quadro 3.4). Você pode também traçar gráficos utilizando o procedimento não interativo do SPSS, mas, para salvar árvores, escrevi isso em um arquivo que está no *site* www.artmed.com.br (**Answers(Chapter 3).pdf**) e no Capítulo 4. Vamos dar uma olhada em um traçado simples: primeiro clique em **Graphs⇒Interactive⇒Bar...** (Gráficos⇒Interativos⇒Barras...) para ativar a caixa de diálogo como a da Figura 3.18. Note que existe uma representação esquemática dos eixos horizontais e verticais e que ambos apresentam um espaço onde é possível arrastar uma variável da lista à esquerda. O eixo y precisa ser a variável dependente ou o que você mensurou, ou simplesmente a coisa para a qual você quer apresentar a média. Nesse caso, será a *arousal* (**excitação**), assim, selecione-a da lista e arraste-a para o espaço que aparece no eixo vertical. O eixo × deve ser a variável na qual queremos dividir os dados da excitação. Para simplificar, vamos apenas traçar as médias para os dois filmes. Selecione a variável *film* (**filme**) da lista à esquerda e arraste-a para o espaço existente no eixo horizontal.[8] Na caixa de diálogo embaixo existe uma seção denominada *Bars Represent...* (Barras representam) e ela ficará ativa somente depois que você especificar a variável do eixo vertical (ela é intitulada *Bars Represent Arousal* (Barras representam excitação) porque especificamos a excitação para o eixo vertical). Por padrão, as barras irão representar as médias (geralmente o que desejamos), mas você pode selecionar a partir da lista suspensa clicando em qualquer outra medida descritiva desde as de tendência central (como a mediana e a moda), medidas de dispersão (variância e desvio padrão) e medidas como a assimetria e a curtose.

Se quisermos adicionar as barras de erros (veja a Seção 7.3), devemos clicar na guia (orelha) *Error Bars* (Barras de Erros) no topo da caixa de diálogo e selecionar Display Error Bars. Por omissão, o SPSS irá mostrar um intervalo de confiança para a média (veja a Seção 1.6.2), mas se você escolher pela lista suspensa clicando em próximo de onde diz *Units* (Unidades), poderá escolher traçar o desvio padrão (veja a Seção 1.4.1) ou o erro padrão da média (veja a Seção 1.6). Finalmente, por padrão, você pode escolher uma representação diferente selecionando um dos quatros botões que estão embaixo. Prefiro o último que apresenta as linhas apontando para fora do topo de cada barra, assim, geralmente seleciono esse.

[8] Para este trabalho é importante que você tenha especificado a variável como nominal (veja a Seção 2.4.3)

Material com direitos autorais

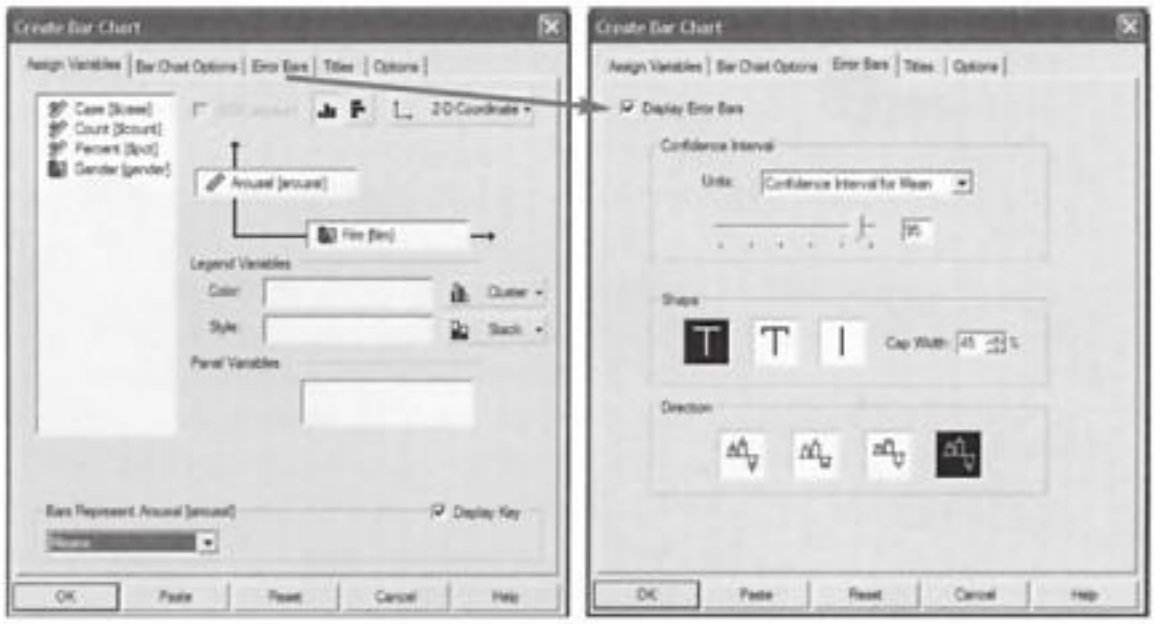


Figura 3.18 Caixas de diálogo para digramas de barras interativo com barras de erros.

A Figura 3.19 mostra o diagrama de barras resultante. Ele simplesmente apresenta as médias (e o intervalo de confiança dessas médias). Esse gráfico demonstra que, em média, as pessoas ficaram mais excitadas pelo filme *Amnésia* do que por *O Diário de Bridget Jones*. Entretanto, inicialmente queríamos verificar os efeitos do gênero, dessa forma, esse gráfico não está informando o que precisamos saber. O gráfico que precisamos é o agrupado (*clustered graph*) e ele é facilmente criado utilizando a mesma caixa de diálogo usada antes (veja a Figura 3.18), mas agora especificando uma segunda variável (nesse caso, gênero) no quadro que diz *Legend Variable* (Variável de Legenda) próximo de onde diz *Color* (Cores). Arraste a variável **gender** (gênero) para esse espaço. A caixa de diálogo final é mostrada na

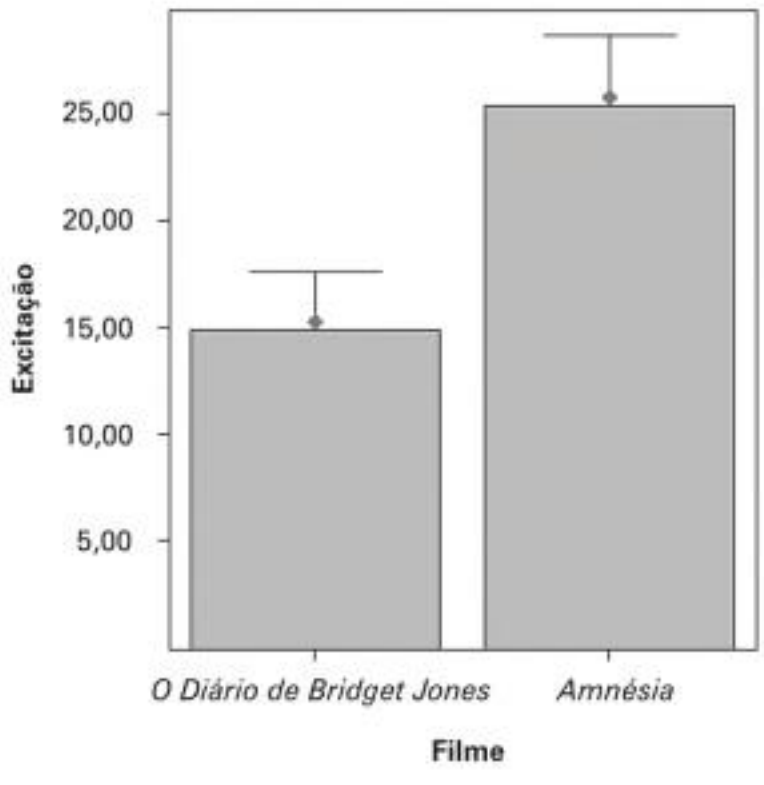


Figura 3.19 Diagrama de barras para a excitação média para cada um dos dois filmes.

Material com direitos autorais

Figura 3.20 e o diagrama de barras resultante, na Figura 3.21. Esse gráfico é mais informativo. Ele mostra o mesmo que o último gráfico: isto é, que a excitação foi mais alta para *Amnésia* do que para *O Diário de Bridget Jones*, mas ele também divide isso por gênero. Olhe primeiro para a excitação média de *O Diário de Bridget Jones*; ele mostra que os homens ficaram de fato mais excitados do que as mulheres. Isso indica que eles gostaram mais do

Figura 3.20 Caixa de diálogo completa para diagrama de barras agrupado.

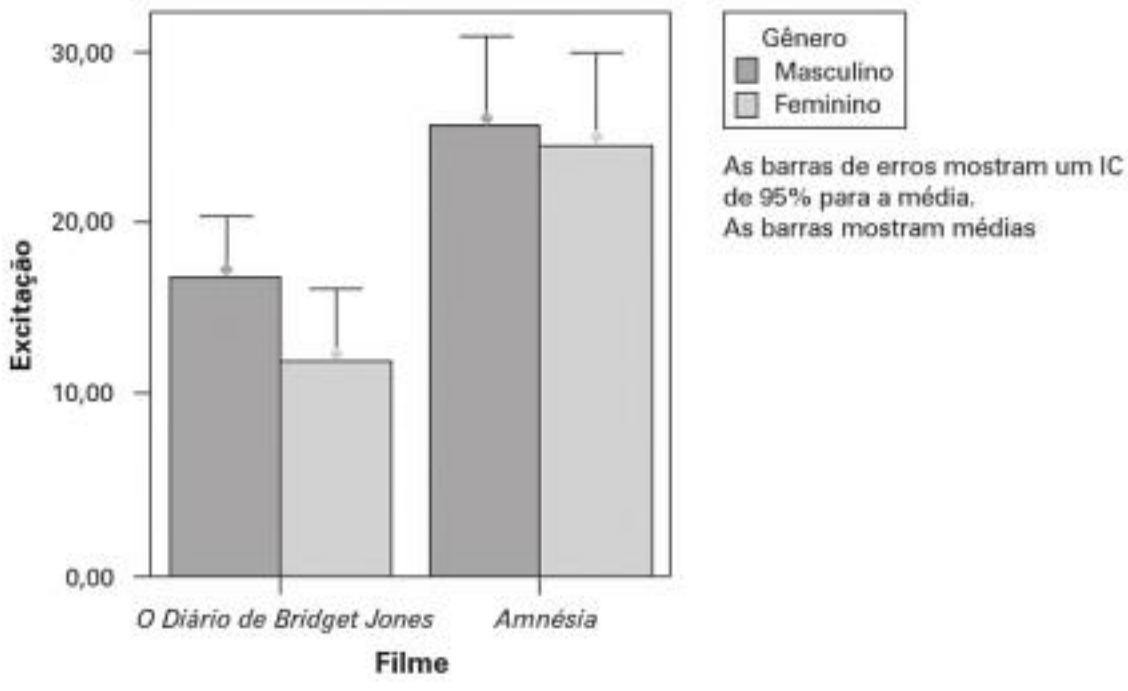


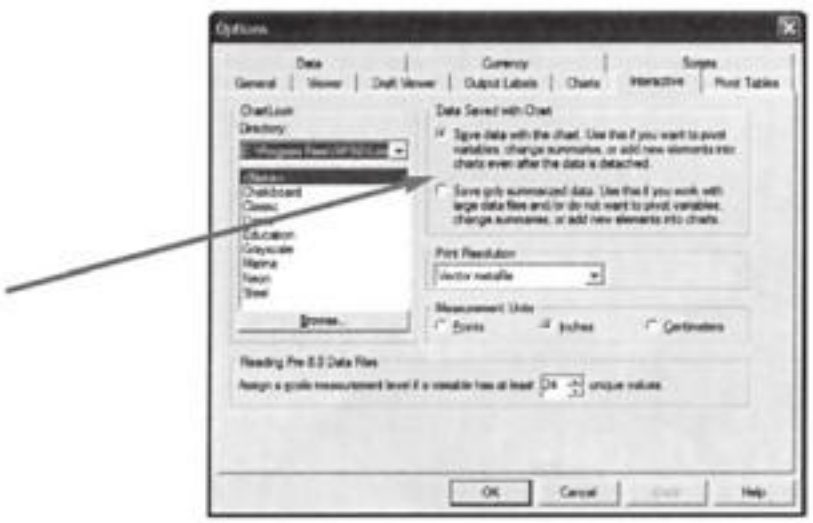
Figura 3.21 Gráfico agrupado mostrando a excitação de homens e mulheres medida durante os dois filmes.

Material com direitos autorais

Quadro 3.4

Transformando gráficos interativos em interativos! ①

A principal vantagem de utilizar um gráfico interativo no SPSS (em vez de um não-interativo) é que você pode atualizá-lo mesmo quando o arquivo de dados estiver fechado. No entanto, para que isso aconteça, você precisa marcar essa capacidade interativa que está desligada. Utilize o menu **Edit⇒Options...** (Editar⇒Opções...) para ter acesso à caixa de diálogo *options* (opções). Dentro dessa caixa de diálogo existe uma guia (orelha)no topo denominada *Interactive* (Interativo), assim, clique nessa orelha e selecione a opção *Save data with the chart* (Salve os dados com o gráfico). Isso transformará seu gráfico interativo em, de fato, interativo! Se você estiver utilizando uma grande base de dados talvez ache melhor desligar novamente essa opção e atualizar o gráfico manualmente se os dados forem alterados ou atualizados.



filme que as mulheres! Contraste isso com o filme *Amnésia*, no qual a excitação entre homens e mulheres é semelhante. Em face disso, a idéia do "filme de garotas" não procede: de fato, parece que homens gostam mais de filmes do tipo "de garotas" do que as próprias garotas (provavelmente por que é a única ajuda que temos para entender o complexo funcionamento da mente feminina!).

Você pode editar quase todos os aspectos de um gráfico clicando duas vezes no gráfico e então clicando na característica que deseja alterar. Você pode trocar as cores das barras, os títulos dos eixos, a escala de cada eixo e assim por diante. Pode, também, tornar as barras tridimensicionais. Contudo, por mais tentador que isso possa ser (ele pode parecer muito bonito) tente resistir a tentação de construir gráficos 3-D (a menos que você esteja realmente representando três variáveis): a dimensão extra não informa de útil e normalmente torna apenas o gráfico mais difícil de interpretar (veja Wright 2003).

3.8 O QUE DESCOBRIMOS SOBRE ESTATÍSTICA? ①

Este capítulo abordou muitas coisas rapidamente! Tentei lhe mostrar a importância de examinar os dados de forma apropriada antes de começar a análise estatística. Primeiro, verificamos a distribuição dos dados e tentamos descobrir possíveis erros que tenham sido cometidos na entrada ou digitação dos dados e também procuramos valores atípicos (*outliers*). Para distribuições não-normais, vimos como transformar os dados a fim de corrigir o problema e como tratar os valores atípicos.

Material com direitos autorais

Se você estiver trabalhando com dados que comparam diferentes grupos, deve examinar as distribuições dentro dos grupos. Você pode também aplicar testes estatísticos para verificar a normalidade dos dados (o teste de Kolmogorov-Smirnov) e olhar para outros gráficos como o diagrama *Q-Q*. Também descobrimos alguns usos para a função *transform* (transformar) do SPSS e para o comando *split file* (dividir arquivo). Finalmente, analisamos rapidamente como traçar médias utilizando gráficos interativos.

3.9 TERMOS-CHAVE QUE DESCOBRIMOS

- Bimodal
- Dados normalmente distribuídos
- Diagramas de caixa e bigodes
- Homogeneidade da variância
- Independência
- Intervalo interquartílico
- Mediana
- Multimodal
- Razão de variâncias
- Teste de Kolmogorov-Smirnov
- Teste de Levene
- Teste de Shapiro-Wilk
- Teste paramétrico
- Transformação
- Valor atípico

3.10 TAREFAS DO ALEX ESPERTO

- Utilizando os dados do Capítulo 2 (que você deve ter salvo, mas se não o fez, copie-os novamente da Tabela 2.1), faça alguns gráficos do número médio de amigos, consumo de álcool, renda e neurose dos estudantes e dos professores. Quem tem mais amigos, bebe mais, ganha mais e é mais neurótico? ①
- Utilizando o arquivo **ChicFlick.sav**, verifique se as distribuições para os dois filmes (ignore o gênero) são normalmente distribuídas. ①
- Utilizando os dados do arquivo **SPSSExam.sav** (lembre que os escores da numerácia pareciam ter assimetria positiva – veja a Figura 3.10), transforme esses dados utilizando uma das transformações descritas no Capítulo; os dados se tornaram normais?

Algumas respostas breves podem ser encontradas no arquivo **Answers(Chapter 3).pdf** no *site* www.artmed.com.br.

3.11 LEITURAS COMPLEMENTARES

TABACHNICK, B. G., FIDELL, L. S. *Using multivariate statistics*. Boston: Allyn & Bacon, 2001 (4ª ed.). O Capítulo 4 é o guia definitivo para a análise de dados!

WRIGHT, D. B. *First steps in statistics*. London: Sage, 2002. O Capítulo 2 é uma boa introdução à construção de gráficos.

[1] IC = Intervalo de confiança

Material com direitos autorais

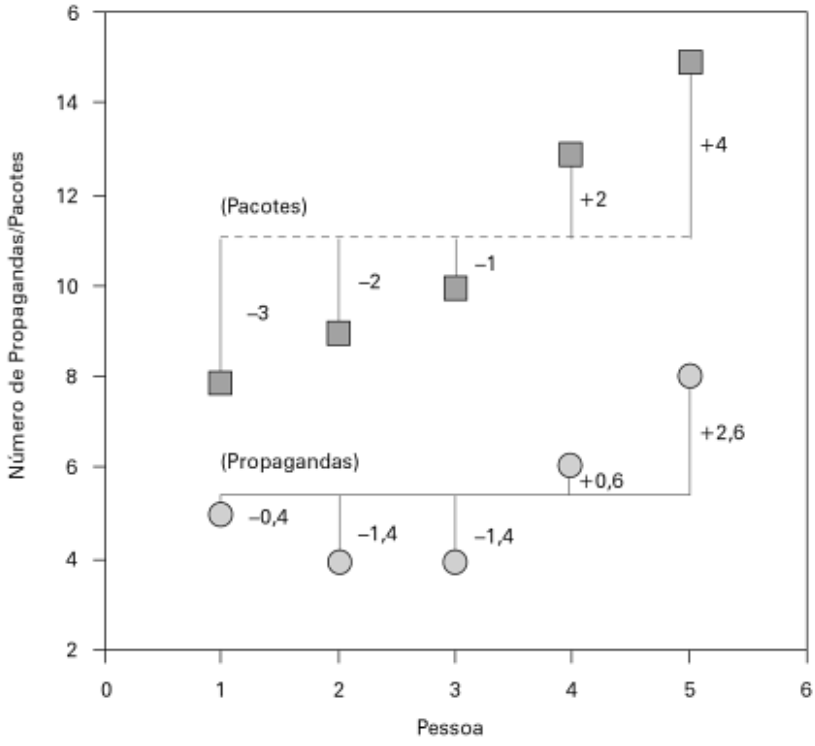


Figura 4.1 Gráfico mostrando as diferenças entre os dados observados e as médias de duas variáveis.

Para os dados da Tabela 4.1 e da Figura 4.1, temos o seguinte valor:

$$\begin{aligned}\text{cov}(x, y) &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1} \\ &= \frac{(-0{,}4)(-3)+(-1{,}4)(-2)}{4} \\ &\quad \frac{+(-1{,}4)(-1)+(-0{,}6)(2)+(2{,}6)+(4)}{4} \\ &= \frac{1{,}2+2{,}8+1{,}4+1{,}2+10{,}4}{4} \\ &= \frac{17}{4} \\ &= 4{,}25\end{aligned}$$

Calcular a covariância é uma boa maneira de avaliar se duas variáveis estão relacionadas entre si. Uma covariância positiva indica que quando uma variável se desvia da média, a outra variável se desvia na mesma direção. Por outro lado, uma covariância negativa indica que, enquanto uma variável se desvia da média (por exemplo, aumenta), a outra se desvia da média na direção oposta (por exemplo, diminui).

Existe, entretanto, um problema com a covariância como uma medida do relacionamento entre variáveis: ela depende das escalas utilizadas para a medida. Assim, covariância não é uma medida padronizada. Por exemplo, se usarmos os dados acima e assumirmos que eles representam duas variáveis mensuradas em quilômetros, então a covariância é de 4,25 (calculada acima). Se convertermos esses dados em milhas (multiplicando todos os valores por 1,609) e calcularmos a covariância novamente, devemos descobrir que ela aumenta para 11. Essa dependência da escala de medidas é um problema porque significa que não podemos comparar a covariância

Material com direitos autorais

de uma maneira objetiva – assim, não podemos dizer se a covariância é particularmente grande ou pequena em relação a outro conjunto de dados a não ser que ambos os conjuntos fossem mensurados nas mesmas unidades.

4.2.2 Padronização e o coeficiente de correlação ①

Para superar o problema da dependência da escala de mensuração, precisamos converter a covariância em um conjunto padrão de unidades. Esse processo é conhecido como ***padronização***. Uma forma básica de padronização seria insistir que em todos os experimentos fossem utilizadas as mesmas unidades de mensuração, digamos, metros – dessa maneira, todos os resultados poderiam ser facilmente comparados. Entretanto, se você fosse medir atitudes, por exemplo, seria muito difícil fazê-lo em metros! Portanto, precisamos de uma unidade de medida na qual qualquer escala de mensuração possa ser convertida. A unidade de medida que usamos é o *desvio padrão*. Falamos sobre essa medida na Seção 1.4.1 e vimos que, como a variância, ela é uma medida da média dos desvios a partir da média. Se dividirmos qualquer distância a partir da média pelo desvio padrão, isso nos dará essa distância em unidades de desvios padrão. Por exemplo, para os dados na Tabela 4.1, o desvio padrão para o número de pacotes comprados é aproximadamente 3,0 (o valor exato é 2,91). Na Figura 4.1, podemos ver que os valores observados para o participante 1 foram três pacotes a menos do que a média (assim, houve um erro de -3 pacotes de balas). Se dividirmos essa diferença, -3, pelo desvio padrão, aproximadamente 3, teremos um valor de -1. Isso nos mostra que a diferença entre o escore do participante 1 e a média foi de -1 desvio padrão. Assim, podemos expressar o desvio a partir da média para um participante em unidades padronizadas dividindo o desvio observado pelo desvio padrão.

Seguindo essa lógica, se quisermos expressar a covariância numa unidade padrão de medida, podemos simplesmente dividi-la pelo desvio padrão. Entretanto, existem duas variáveis e, dessa maneira, dois desvios padrão. Quando calculamos a covariância, na verdade calculamos dois desvios (um para cada variável) e, então, os multiplicamos. Por conseguinte, fazemos o mesmo com os desvios padrão: nós os multiplicamos e dividimos pelo produto dessa multiplicação. A covariância padronizada é conhecida como *coeficiente de correlação*. E é definida pela equação (4.3) na qual s_x é o desvio padrão da primeira variável e s_y é o desvio padrão da segunda variável (todas as outras letras são as mesmas da equação definindo a covariância):

$$r = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} \quad (4.3)$$

O coeficiente na equação (4.3) é conhecido como o ***coeficiente de correlação* produto-momento de Pearson** (ou **simplesmente coeficiente de correlação de Pearson**) e foi inventado por Pearson (que surpresa!).[1] Se olharmos para a Tabela 4.1, vemos que o desvio padrão para o número de propagandas assistidas (s_x) foi 1,67 e para o número de pacotes de balas compradas (s_y) foi 2,92. Se multiplicarmos esses valores, temos $1{,}67 \times 2{,}92 = 4{,}88$. Agora, tudo o que precisamos fazer é pegar a covariância, que calculamos algumas páginas atrás como sendo 4,25, e dividir por esse produto. Isso nos dá $r = 4{,}25/4{,}88 = 0{,}87$ (esse é o mesmo valor que você encontrará mais tarde na Saída 4.1 do SPSS).

Padronizando a covariância, encontramos um valor que deve estar entre -1 e $+1$ (se você encontrar um coeficiente de correlação menor de que -1 ou maior do que $+1$, pode estar certo que alguma coisa saiu errada!) Um coeficiente de $+1$ indica que as duas variáveis são perfeitamente correlacionadas de forma positiva, assim, enquanto uma variável aumenta, a outra

[1] Você irá encontrar o coeficiente de correlação momento-produto de Pearson representado por ambos r e R. Tipicamente, a letra maiúscula é usada num contexto de regressão porque ela representa o coeficiente de correlação múltiplo; entretanto por alguma razão, quando elevamos r ao quadrado (como na Seção 4.5.3), é utilizado o R com letra maiúscula. Não me pergunte o porquê – é só para me confundir, eu acho.

Material com direitos autorais

Dica da Samanta Ferrinho

- Uma medida grosseira do relacionamento entre variáveis é a *covariância*.
- Se padronizarmos esse valor temos *o coeficiente de correlação de Pearson, r.*
- O coeficiente de correlação deve variar entre −1 e +1.
- Um coeficiente de +1 indica um relacionamento positivo perfeito, um coeficiente de −1 indica um relacionamento negativo perfeito, um coeficiente 0 indica que não existe relacionamento linear.
- O coeficiente de correlação é uma medida do tamanho de um efeito comumente utilizada: valores de ± 0,1 representam um efeito pequeno, ± 0,3 representa um efeito médio e ± 0,5, um efeito grande.

aumenta proporcionalmente. Inversamente, um coeficiente de −1 indica um relacionamento negativo perfeito: se uma variável aumenta, a outra diminui por um valor proporcional. Um coeficiente 0 indica ausência de relacionamento linear, isto é, se uma variável muda, a outra permanece praticamente igual. Também vimos na Seção 1.8.4 que, como o coeficiente de correlação é uma medida padronizada de um efeito observado, ele é uma medida comumente usada do tamanho de efeito, e que valores de ± 0,1 representam um efeito pequeno, ± 0,3, um efeito médio e ± 0,5, um efeito grande.

4.3 ENTRADA DE DADOS PARA REALIZAR UMA ANÁLISE DE CORRELAÇÃO NO SPSS ①

A entrada de dados para a correlação, a regressão e a regressão múltipla é simples porque cada variável é introduzida em uma coluna separada. Portanto, para cada variável que você mediu, crie uma variável no editor de dados com um nome apropriado, e digite os escores de um participante numa linha do editor de dados. Poderá ocorrer que você tenha uma ou mais variáveis categóricas (como o gênero) e essas variáveis também podem ser digitadas numa coluna (mas lembre-se de definir o valor apropriado dos rótulos). Por exemplo, se quiséssemos calcular a correlação entre as duas variáveis na Tabela 4.1, deveríamos entrar esses dados como na Figura 4.2. Ao longo deste capítulo iremos analisar um conjunto de dados baseado no desempenho em uma prova da faculdade. Os dados para vários exemplos estão armazenados num único arquivo disponível no *site* www.artmed.com.br chamado de **ExamAnxiety.sav.** Se você abrir esse arquivo, verá que esses dados serão exibidos no editor em colunas separadas e que **gender** (gênero) foi codificado de forma correta. Iremos descobrir a que cada uma das variáveis se refere neste capítulo.

4.4 REPRESENTANDO RELACIONAMENTOS GRAFICAMENTE: O DIAGRAMA DE DISPERSÃO ①

Antes de conduzir qualquer análise de correlação é *essencial* traçar um diagrama de dispersão para ver a tendência geral dos dados.

	adverts	packets
1	5.00	8.00
2	4.00	9.00
3	4.00	10.00
4	6.00	13.00
5	8.00	15.00
6		

Figura 4.2 Entrada de dados para a correlação. O editor de dados mostra que o participante um viu cinco propagandas (**adverts**) e subsequentemente comprou oito pacotes (**packets**) de balas.

Material com direitos autorais

Um diagrama de dispersão é simplesmente um gráfico que coloca o escore de cada pessoa em uma variável contra seu escore em outra (e seu escore numa terceira variável pode ser incluído em um diagrama de dispersão 3D). Um diagrama de dispersão nos diz muitas coisas sobre os dados, por exemplo, se parece haver um relacionamento entre as variáveis, que tipo de relacionamento ele é e se quaisquer casos são significativamente diferentes dos outros. Um caso que difere substancialmente da tendência geral dos dados é conhecido como um *valor atípico* e tais casos podem distorcer severamente o coeficiente de correlação (veja o Quadro 3.1 e a Seção 5.6.1.1 para mais detalhes). Podemos usar um diagrama de dispersão para mostrar se algum caso parece um valor atípico.

Traçar um diagrama de dispersão usando o SPSS é fácil demais. Você pode traçá-lo usando o menu dos gráficos interativos ou usando gráficos não-interativos. Vamos examinar os dois tipos de gráficos porque há diagramas de dispersão que você pode criar apenas com gráficos não-interativos.

4.4.1 Diagrama de dispersão simples ①

Esse tipo de diagrama de dispersão serve para relacionar apenas duas variáveis. Por exemplo, uma psicóloga estava interessada nos efeitos do estresse pré-prova no desempenho na prova. Assim, ela planejou e validou um questionário para estimar o estado de ansiedade relacionado às provas (chamado de Questionário de Ansiedade Pré-Prova, ou QAP). Esse instrumento produziu uma medida de ansiedade numa escala de 1 a 100. A ansiedade foi medida antes de uma prova e a percentagem obtida para cada aluno foi usada para avaliar o desempenho na prova. Antes de verificar se essas variáveis estavam correlacionadas, a psicóloga fez um diagrama de dispersão das variáveis envolvidas (os dados estão no arquivo **ExamAnxiety.sav** que você já deve ter aberto no SPSS). Para representar graficamente essas duas variáveis, você pode usar um gráfico interativo. Para tanto, selecione **Graphs⇒Interactive⇒Scatterplot...** (Gráficos⇒Interativo⇒Diagrama de Dispersão...) para abrir uma caixa de diálogo como a da Figura 4.3. Nessa caixa de diálogo, todas as variáveis do editor de dados são mostradas no lado esquerdo com espaços vazios no lado direito. Você simplesmente clica numa variável da lista à esquerda e a arrasta para o local apropriado. Em qualquer gráfico o eixo vertical é conhecido como o eixo *y* (ou ordenada) do gráfico. Essa variável deve ser a dependente (o resultado que foi medido)[2],que nesse caso é **exam performance (%) [exam]** (desempenho da prova (%) [prova]). O eixo horizontal é conhecido como o eixo × (ou abscissa) do gráfico. Essa variável deveria ser a variável independente, que é o caso em **exam anxiety [anxiety]** (ansiedade pré-prova [ansiedade]). A caixa de diálogo da Figura 4.3 tem uma representação esquemática do eixo *y* (a seta apontando para cima) e do eixo *x* (a seta apontando para a direita) e cada uma tem um espaço. Para especificar essas variáveis, precisamos arrastá-las para o espaço apropriado. Assim, arraste **exam performance (%) [exam]** (desempenho na prova (%) [prova]) para o espaço do eixo *y* e arraste **exam anxiety [anxiety]** (ansiedade pré-prova [ansiedade]) para o espaço do eixo *x*.

Mais abaixo na caixa de diálogo existem outros espaços nos quais você pode arrastar as variáveis. Você pode usar uma variável de grupo para definir diferentes categorias no diagrama de dispersão (ele irá exibir cada

[2] Na pesquisa experimental, é comum representar a variável independente no eixo horizontal e a variável dependente no eixo vertical. Nessa forma de pesquisa controlada, a implicação é que mudanças na variável independente (a variável que o pesquisador manipulou) causam mudanças na variável dependente. Na pesquisa correlacional, variáveis são mensuradas simultaneamente e, assim, nenhum relacionamento de causa e efeito pode ser estabelecido. Como tal, esses termos não devem ser encarados de forma rígida!

Material com direitos autorais

Figura 4.3 Caixa de diálogo para um diagrama de dispersão simples.

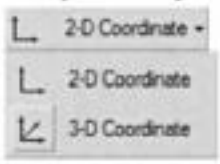
categoria usando uma cor diferente, símbolo ou gráfico). Essa função é útil para ver o relacionamento entre duas variáveis para grupos de pessoas. No exemplo atual, temos dados identificando se o estudante é do sexo masculino ou feminino, assim, podemos usar essas opções da seguinte maneira:

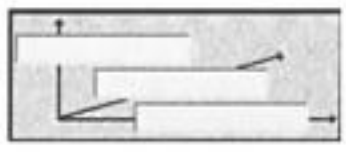
- **Color** (Cor): Se você selecionar **Gender [gender]** (Gênero [gênero]) da lista e arrastá-lo para esse espaço, o diagrama de dispersão resultante exibirá os dados masculinos numa cor diferente dos dados femininos.
- **Style** (Estilo)**:** Se você selecionar **Gender [gender]** (Gênero [gênero]) da lista e arrastá-lo para esse espaço, o diagrama de dispersão resultante usará símbolos diferentes para os dados masculinos e femininos. Isso é o que escolhemos fazer e a caixa de diálogo resultante deverá ser parecida com a da Figura 4.3.
- **Size** (Tamanho): Se você selecionar **Gender [gender]** (Gênero [gênero]) da lista e arrastá-lo para esse espaço, os símbolos dos dados para o sexo masculino e feminino ficarão de tamanhos diferentes no diagrama de dispersão resultante.
- **Panel Variables** (Variáveis do Painel)**:** Se você selecionar **Gender [gender]** (Gênero [gênero]) da lista e arrastá-lo para esse espaço, o SPSS irá produzir dois diagramas de dispersão (lado a lado) – um para o sexo masculino e outro para o feminino.
- **Label Cases by** (Rotular Casos por)**:** Se você selecionar **Gender [gender]** (Gênero [gênero]) da lista e arrastá-lo para esse espaço, o SPSS colocará um rótulo próximo a cada ponto de dados no diagrama de dispersão indicando se aqueles dados são do sexo masculino ou feminino. Com muitos pontos de dados isso irá dificultar a leitura! Um dos usos pode ser se você tiver uma variável com o nome de um participante, ou um código; nesse casso, cada ponto no diagrama de dispersão será rotulado com um código que identifica a pessoa, o que pode ser útil para identificar valores atípicos e, então, ser realmente capaz de descobrir qual caso é um valor atípico!

No topo da caixa de diálogo há várias orelhas que dão acesso a outras opções. Se você clicar em *Titles* (Títulos), por exemplo, verá uma caixa de diálogo que dá um espaço para digitar um título para o diagrama de dispersão. Uma orelha que pode ser útil (especialmente quando você ler o próximo capítulo) é a *Fit* (Ajustar) (Figura 4.4). Dentro dessa caixa de diálogo você pode escolher uma linha para traçar selecionando um *Method* (Método) da lista suspensa. Se você escolher *regression* (regressão), o SPSS se ajustará à linha reta que melhor representa o relacionamento entre as variáveis no seu diagrama de dispersão (veja o Capítulo 5). Se você escolher *mean* (média), o SPSS traça uma linha horizontal representando a média, e se você escolher *smoother* (suavizar), uma linha curva que melhor descreva o relacionamento entre as variáveis no seu diagrama de dispersão será ajustada. Se você selecionar ☑ Total, o SPSS traçará essas linhas para todos os dados, e se você se-

Material com direitos autorais

brilhante)[3]. Por exemplo, imagine que a sua pesquisadora decidiu que ansiedade pré-prova pode não ser o único fator contribuindo para o desempenho na prova. Então, ela também pediu para os participantes manterem um diário de revisão, e ela calculou o número de horas gastas revisando a matéria da prova. A pesquisadora queria ver o relacionamento entre essas variáveis simultaneamente. Para criar um diagrama de dispersão em 3D, use uma caixa de diálogo como a da Figura 4.3. Nessa caixa de diálogo existe uma seta com um menu suspenso rotulado de *2-D Co-ordinate* (Coordenadas 2D). Clique na seta para acessar o menu suspenso e você verá uma segunda opção rotulada de *3-D Co-ordinate* (Coordenadas 3D). Clique nessa opção e uma representação esquemática dos eixos irá mudar para incluir o terceiro eixo: o eixo *z*. Usando a Figura 4.3 como um guia, complete a caixa de diálogo como anteriormente (tente colocar a variável do gênero na caixa colorida dessa vez), mas selecione *3-D Co-ordinate* (Coordenada 3D) e então clique e arraste a variável **Time spent revising [revise]** (Tempo gasto revisando [revisão]) para o espaço do eixo *z*. O diagrama de dispersão resultante é mostrado na Figura 4.6. Geralmente, os diagramas de dispersão em 3D podem ser difíceis de interpretar e, assim, sua utilidade em explorar dados pode ser limitada; entretanto, eles são ótimos para exibir dados de maneira concisa. Agora, se você clicar duas vezes nesse gráfico, você será capaz de girar os eixos para ter uma idéia melhor do relacionamento entre as três variáveis – isso será útil para você; contudo, você será obrigado a decidir que ângulo melhor representa o relacionamento quando quiser colocar o gráfico no papel em 2D! Divirta-se com o gráfico e veja se você consegue a mesma figura que eu (a tela próxima desse texto irá ajudar!)





4.4.3 Diagrama de dispersão sobreposto ①

Existem dois tipos de diagramas de dispersão que você não pode fazer usando gráficos interativos: o diagrama de dispersão sobreposto e o diagrama de dispersão matricial. Ambos são muito úteis e vale a pena descobrir o porquê. Um diagrama de dispersão sobreposto é aquele em que muitos pares de variáveis são traçados nos mesmos eixos. Imagine que nossa pesquisadora quisesse verificar o papel da ansiedade pré-prova e do tempo de revisão no desempenho da prova (mas não o relacionamento entre a ansiedade pré-prova e revisão); seria útil traçar os relacionamentos entre tempo de revisão e desempenho na prova e entre ansiedade pré-prova e desempenho simultaneamente. Para tirar o melhor proveito desse tipo de diagrama de dispersão, mantenha uma variável constante e trace-a contra outras. No exemplo da prova, no qual o efeito da ansiedade e do tempo de revisão no desempenho da prova é relevante, você deveria traçar ansiedade **(anxiety)** (*X*) contra prova (**exam**) (*Y*) e, então, sobrepor revisão (**revise**) (*X*) contra desempenho na prova **(exam)** (*Y*).

Para traçar essas combinações, simplesmente use os menus como segue: **Graphs⇒Scatter...** (Gráficos⇒Espalhamento...). Isso ativa a caixa de diálogo da Figura 4.7, que fornece quatro opções para diferentes tipos de diagramas de dispersão disponíveis. Por omissão, um diagrama simples de dispersão é selecionado como é mostrado pela borda preta em volta da figura. Se você deseja traçar um diagrama de dispersão diferente, mova o cursor sobre as outras figuras e clique com o botão esquerdo do mouse. Depois de selecionar o diagrama de dispersão, clique em Define (Definir).

[3] Não é, na verdade, errado usar um diagrama em 3D quando você está representando somente duas variáveis. Quero dizer que nada de ruim irá lhe acontecer (a não ser que você o mostre para certos estatísticos que o trancarão numa sala e farão você escrever 75172 vezes no quadro "eu não devo fazer gráficos em 3D"), apenas não é uma maneira clara de apresentar os dados.

Material com direitos autorais

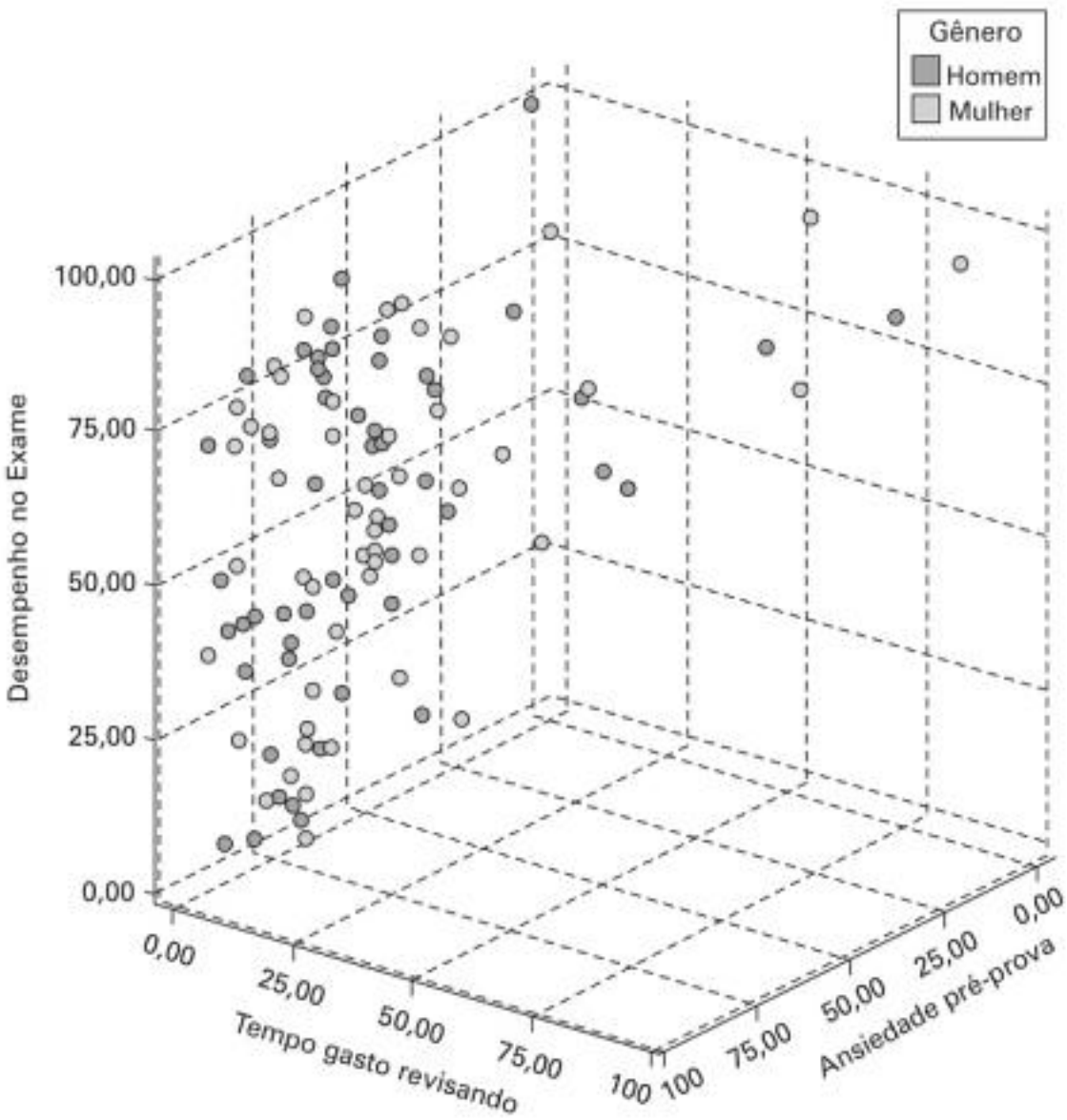


Figura 4.6 Um diagrama de dispersão em 3 D do desempenho da prova traçado contra a ansiedade pré-prova e o tempo gasto revisando.

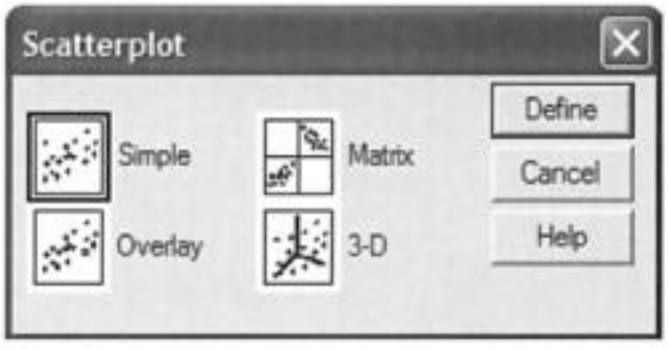


Figura 4.7 Caixa de diálogo do diagrama de dispersão principal.

Para traçar um diagrama de dispersão sobreposto, clique em *Overlay* (Sobreposto) na caixa de diálogo principal do diagrama de dispersão (*scatterplot*) (veja a Figura 4.7) e uma caixa de diálogo como a da Figura 4.8 aparece. Para selecionar um par de variáveis, clique em uma variável da lista (isso irá aparecer como *Variable 1* [Variável 1] na seção rotulada de *Currrent Selections* [Seleções Atuais]) e, então, selecione uma segunda variável (essa estará listada como *Variable* 2 [Variável 2]). Transfira o par de variáveis clicando em ▸. O par irá aparecer no espaço rotulado de *Y–X Pairs* (Pares Y–X). A ordem das variáveis está relacionada com o eixo em que elas estarão traçadas, assim, na Figura 4.8 **exam** (prova) irá aparecer no eixo ***y*** e **anxiety** (ansiedade) no eixo ***x***. O segundo par, **exam** e **revise** (prova e revisão) deve, também, ser transferido para

Material com direitos autorais

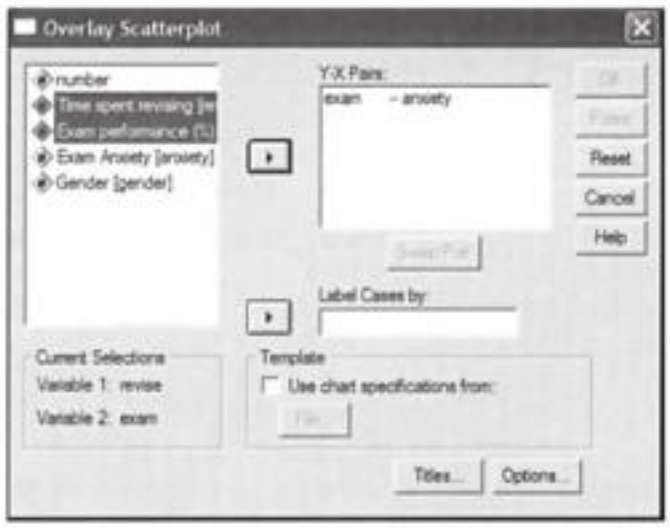


Figura 4.8 Caixa de diálogo para criar um diagrama de dispersão sobreposto. O primeiro par de variáveis já foi especificado e o segundo par foi selecionado, mas não transferido para a lista dos pares.

a lista de pares Y–X, de modo que **exam** (prova) esteja listada em primeiro lugar (e, assim, será traçada no eixo **y**) e **revise** (revisão) em segundo lugar (assim, será traçada no eixo **x**). Se ao transferir duas variáveis, elas não aparecerem da forma desejada, você poderá trocar a ordem das variáveis (e, portanto, o eixo no qual elas serão representadas) clicando em Swap Pair (trocar par).

Fica claro, a partir da Figura 4.9, que, embora a ansiedade esteja negativamente relacionada ao desempenho na prova, parece que a variável desempenho na prova está positivamente relacionada ao tempo de revisão. Assim, enquanto o tempo de revisão aumenta, o desempenho na prova aumenta também, mas, enquanto a ansiedade aumenta, o desempenho na prova diminui. O diagrama de dispersão sobreposto mostra de forma clara esses diferentes relacionamentos.

4.4.4 Diagrama de dispersão matricial ①

Em vez de traçar muitas variáveis nos mesmos eixos em um diagrama 3D (o que pode ser difícil de interpretar com vários pares de variáveis), é possível traçar uma matriz dos diagramas de dispersão. Esse tipo de diagrama permite ver o relacionamento entre todas as combinações dos diferentes pares de va-

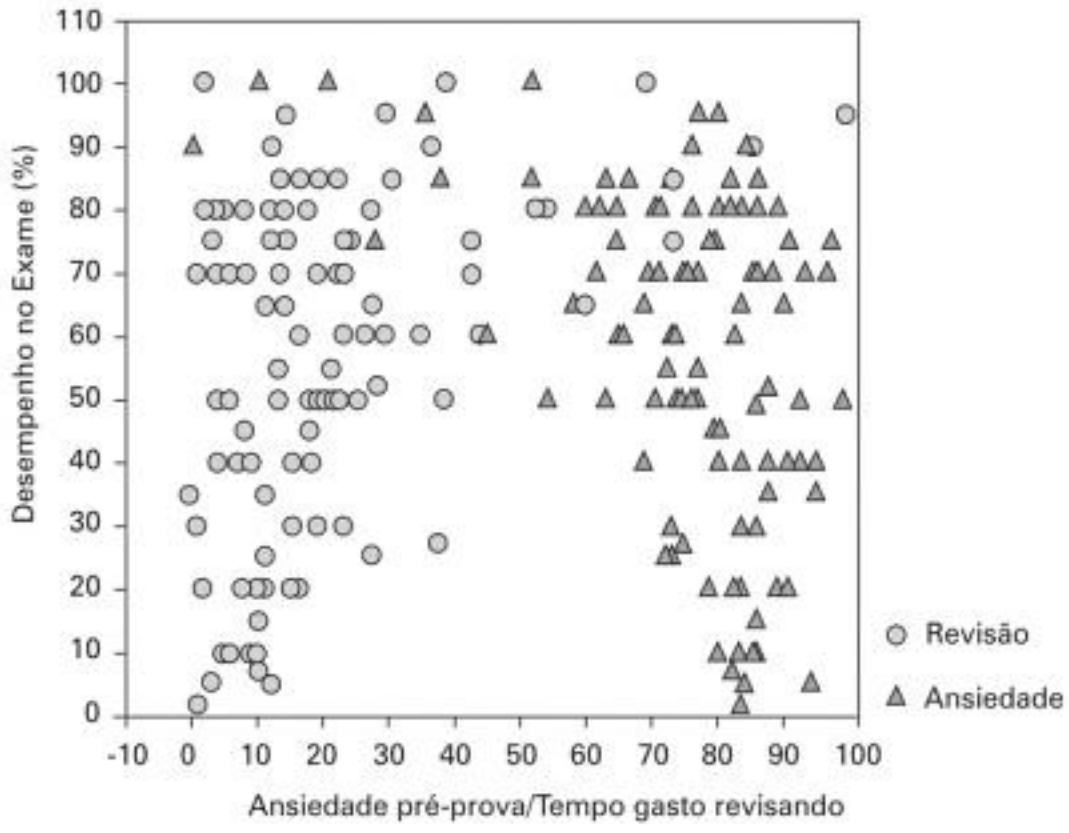


Figura 4.9 Um diagrama de dispersão do desempenho na prova *versus* a ansiedade pré-prova e tempo gasto revisando.

Material com direitos autorais

riáveis. Para obter um diagrama de dispersão matricial para os dados usados no diagrama de dispersão sobreposto, selecione a opção diagrama de dispersão Matricial (*Matrix*) na caixa de diálogo principal do diagrama de dispersão (Figura 4.7). A caixa de diálogo completa é mostrada na Figura 4.10: as variáveis estão listadas à esquerda e qualquer uma que você queira representar deverá ser transferida para a caixa rotulada de *Matrix Variables* (Variáveis Matriciais) usando a tecla . Para nossos dados, selecione e transfira **Exam performance (%) [exam], Exam Anxiety [anxiety]** e **Time spent revising [revise]** (Desempenho na Prova (%) [prova], Ansiedade Pré-Prova [ansiedade] e Tempo gasto revisando [revisão]) para o quadro denominado *Matriz Variables* (Variáveis Matriciais). Como ocorreu com o diagrama de dispersão simples, existe uma opção de separar o diagrama de dispersão por uma variável de grupo especificando *Set Markers by...* (Determine marcadores por), mas não há necessidade de usar essa opção para esses dados. O diagrama de dispersão resultante do desempenho na prova contra ansiedade pré-prova e tempo de revisão é mostrado na Figura 4.11.

Os seis diagramas de dispersão na Figura 4.11 representam as várias combinações de cada variável traçada uma contra a outra.

Figura 4.10 Caixa de diálogo para o diagrama de dispersão matricial.

Assim, a grade de referências representa os seguintes diagramas:

- **B1:** desempenho na prova (Y) *versus* ansiedade (X)
- **C1:** desempenho na prova (Y) *versus* tempo de revisão (X)
- **A2:** ansiedade (Y) *versus* desempenho na prova (X)
- **C2:** ansiedade (Y) *versus* tempo de revisão (X)
- **A3:** tempo de revisão (Y) *versus* desempenho na prova (X)
- **B3:** tempo de revisão (Y) *versus* ansiedade (X)

Dessa maneira, os três diagramas de dispersão abaixo da diagonal da matriz são os mesmos diagramas dos acima da diagonal, mas com eixos ao contrário. Dessa matriz podemos ver que o tempo de revisão e a ansiedade estão inversamente relacionados (assim, quanto mais tempo gasto revisando, menos ansiedade o participante teve na prova). Também, no diagrama de dispersão do tempo de revisão contra ansiedade (grade C2 e B3) parece que há um possível valor atípico – há um único participante que gastou pouquíssimo tempo revisando e ficou pouco ansioso em relação à prova. Como todos os participantes que tiveram baixa ansiedade obtiveram notas altas na prova, podemos deduzir que essa pessoa também foi bem na prova (você não odeia um Alex Esperto!). Poderíamos examinar esse caso mais detalhadamente se achássemos que seu comportamento foi provocado por algum fator externo (como tomar estimulantes para o cérebro!). Os diagramas de dispersão matriciais são muito convenientes para examinar pares de relacionamento entre variáveis. Entretanto, não recomendo traçá-los para mais de três ou quatro variáveis porque eles se tornam muito confusos.

4.5 CORRELAÇÃO BIVARIADA ①

Tendo uma visão preliminar dos dados, podemos agora realizar a análise de correlação. Para tanto, iremos retornar aos dados da

Material com direitos autorais

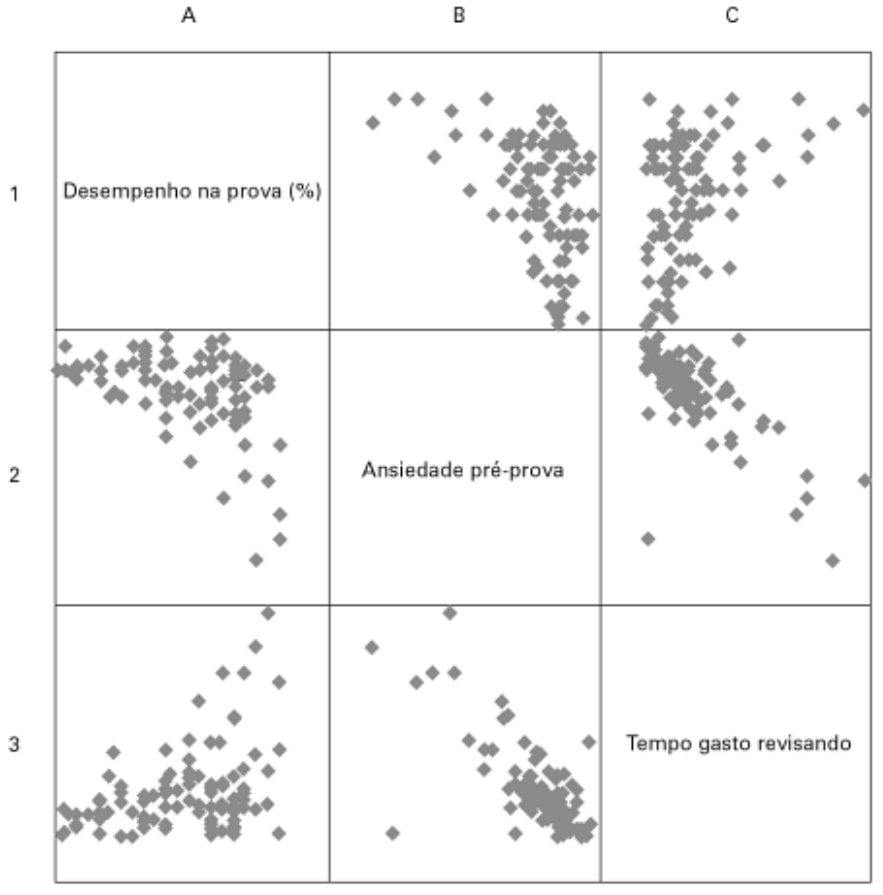


Figura 4.11 O diagrama de dispersão matricial do desempenho na prova, ansiedade pré-prova e tempo de revisão. Referências na grade foram adicionadas para esclarecimentos.

Tabela 4.1, que examinava se havia um relacionamento entre o número de propagandas assistidas e o número de pacotes de balas *toffee* comprados posteriormente. Crie duas colunas no editor de dados e digite esses dados (os preguiçosos podem acessar o arquivo **Advert.sav** no *site* www.artmed.com.br).

Existem dois tipos de correlação: *bivariada* e *parcial*. Uma correlação bivariada é uma correlação entre duas variáveis (como foi descrito no início deste capítulo), enquanto uma correlação parcial determina o relacionamento entre variáveis "controlando" o efeito de uma ou mais variáveis. O coeficiente de correlação momento-produto de Pearson (descrito anteriormente) e o rô de Spearman (veja a Seção 4.5.4) são exemplos de coeficientes de correlação bivariada. Para conduzir uma correlação bivariada, você precisa encontrar a opção *Correlate* (Correlacionar) do menu *Analyse* (Analisar). A caixa de diálogo principal é acessada pelo caminho do menu **Analyse⇒Correlate⇒Bivariate...** (Analisar⇒Correlacionar⇒Bivariada...) e é mostrada na Figura 4.12. Usando a caixa de diálogo é possível selecionar quais das três correlações você quer determinar. O padrão estabelecido é a correlação de Pearson, mas você também pode calcular a de Spearman e a de Kendall – veremos as diferenças entre esses coeficientes

Material com direitos autorais

entre duas variáveis. Entretanto, se você quer determinar se o coeficiente de correlação é significativo, mais suposições são necessárias: para que a estatística teste seja válida os dados devem ser normalmente distribuídos. Embora queiramos que ambas as variáveis sejam distribuídas normalmente, há uma exceção para essa regra: uma das variáveis pode ser categórica contanto que ela tenha somente duas categorias (na realidade, se você olhar a Seção 4.5.6, verá que isso é o mesmo que fazer um teste-*t*, mas estou me apressando um pouco). De qualquer maneira, se os seus dados não são normais (veja o Capítulo 3) ou não são mensurados em intervalos, você não deve selecionar a correlação de Pearson (Figura 4.14).

Figura 4.14 Karl Pearson.

Acesse novamente as dados do arquivo **ExamAnxiety.sav** para que possamos calcular a correlação entre o desempenho na prova, tempo de revisão e ansiedade pré-prova. Os dados do desempenho na prova estudados são paramétricos e, assim, a correlação de Pearson pode ser aplicada. Acesse a caixa de diálogo principal das *bivariate correlations* (correlações bivariadas) (**Analyze⇒Correlate⇒Bivarite...**) (Analisar⇒Correlacionar⇒Bivariada...) e transfira **exam, anxiety** e **revise** (prova, ansiedade e revisar) para o quadro rotulado de *Variables* (Variáveis) (Figura 4.15). A caixa de diálogo também nos permite especificar se o teste é uni ou bilateral. Testes unilaterais devem ser usados quando existe uma direção específica para a hipótese que está sendo testada e testes bilaterais devem ser usados quando um relacionamento é esperado, mas a direção do relacionamento não é previsível. Nosso pesquisador previu que em altos índices de ansiedade o desempenho da prova seria baixo. Portanto, o teste para essas variáveis deve ser unilateral porque antes dos dados serem cole-

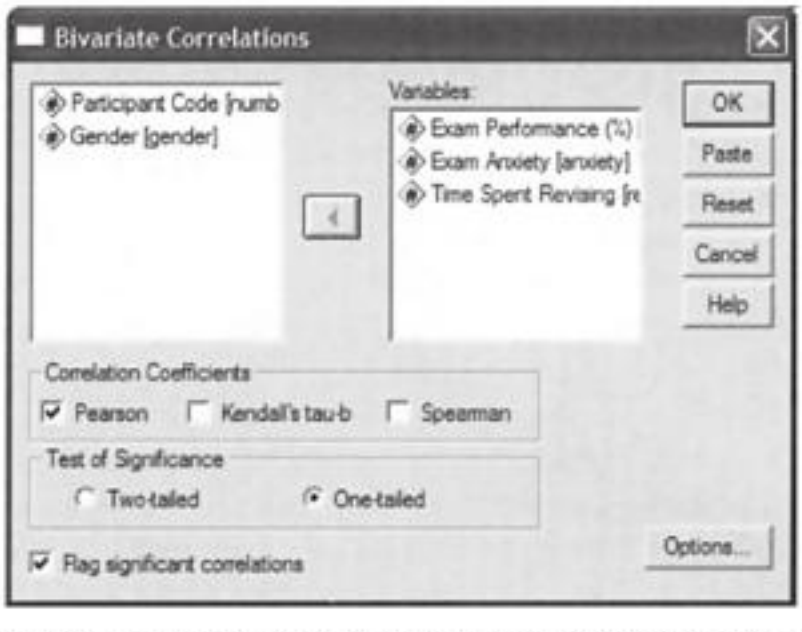


Figura 4.15 Caixa de diálogo completa para os dados do desempenho na prova.

Material com direitos autorais

tados o pesquisador previu um tipo específico de relacionamento. E mais, uma correlação positiva entre tempo de revisão e desempenho na prova foi previsto, assim, esse teste também deve ser unilateral. Para assegurar que iremos obter os valores de significância unilaterais selecione One-tailed (unilateral) e clique em OK.

A saída 4.2 do SPSS fornece uma matriz dos coeficientes de correlação para as três variáveis. Embaixo de cada coeficiente de correlação é apresentado o valor da significância da correlação e o tamanho da amostra (*N*) no qual ele é baseado. Cada variável é perfeitamente correlacionada com ela própria (obviamente) e, assim, $r = 1$ ao longo da diagonal na tabela. O desempenho na prova é negativamente relacionado à ansiedade pré-prova com um coeficiente $r = -0{,}441$ e existe uma probabilidade de menos de 0,001 que esse coeficiente de correlação tenha ocorrido por acaso numa amostra de 103 pessoas (como está indicado pelos dois asteriscos depois do coeficiente). Esse valor da significância nos diz que a probabilidade dessa correlação ser "um golpe de sorte" é muito baixa (próximo a zero, na verdade). Desde já, podemos ter certeza de que o relacionamento entre desempenho na prova e ansiedade é genuíno. Geralmente, cientistas sociais aceitam qualquer valor de probabilidade abaixo de 0,05 como sendo estatisticamente significativo e, assim, tal valor é considerado como um indicativo de um efeito genuíno (o SPSS irá marcar qualquer coeficiente de correlação significativo, nesse nível, com um asterisco). A saída mostra, também, que o desempenho na prova está positivamente relacionado com o tempo gasto revisando, com um coeficiente de $r = 0{,}397$, o qual também é significativo a $p < 0{,}001$. Finalmente, a ansiedade pré-prova parece ser negativamente relacionada ao tempo gasto revisando ($r = -0{,}709$, $p < 0{,}001$).

Em termos psicológicos, isso tudo significa que, à medida que a ansiedade sobre uma prova aumenta, a percentagem de acertos obtida naquele exame diminui. Inversamente, à medida que o tempo de revisão aumenta, a percentagem obtida na prova aumenta. Finalmente, à medida que o tempo de revisão aumenta, a ansiedade do estudante sobre a prova diminui. Assim, há um inter-relacionamento complexo entre as três variáveis.

4.5.2 Um alerta sobre interpretação: Causalidade ①

Devemos ser cuidados ao interpretar coeficientes de correlação porque eles não dão indicação da direção da *causalidade*. Assim,

Saída 4.2 do SPSS Saída do SPSS para a correlação de Pearson

***Correlations* (Correlações)**

		Exam performance (%) (Desempenho na prova)	*Exam Anxiety* (Ansiedade pré-prova)	*Time spent revising* (Tempo gasto revisando)
Exam performance (%) (Desempenho na prova)	*Pearson Correlation* (Correlação de Pearson)	1.000	− 0.441**	0.397**
	Sig. (1 tailed) (Sig. unilateral)	.	0.000	0.00
	N	103	103	103
Exam Anxiety (Ansiedade pré-prova)	*Pearson Correlation* (Correlação de Pearson)	− 0.441**	1.000	− 0.709**
	Sig. (1 tailed) (Sig. unilateral)	0.000	.	0.000
	N	103	103	103
Time spent revising (Tempo gasto revisando)	*Pearson Correlation* (Correlação de Pearson)	0.397**	− 0.709**	1.000
	Sig. (1 tailed) (Sig. unilateral)	0.000	0.000	.
	N	103	103	103

Correlation is significant at the .01 level (1-tailed) (Correlação é significativa ao nível de 0,01 (unilateral))

Material com direitos autorais

em nosso exemplo, embora possamos concluir que o desempenho na prova diminui à medida que a ansiedade pré-prova aumenta, não podemos dizer que uma alta ansiedade pré-prova *causa* péssimo desempenho na prova. Essa precaução tem dois motivos:

- **O problema da terceira variável:** Em qualquer correlação bivariada a causalidade entre duas variáveis não pode ser dada por certo, porque podem ter outras variáveis, medidas ou não, afetando os resultados. Isso é conhecido como o problema da *terceira variável* ou o *tertium quid*. No nosso exemplo, você pode ver que o tempo de revisão se relaciona significativamente a ambos, desempenho na prova e ansiedade pré-prova, e não há maneira de dizer qual das duas variáveis independentes, se ambas, causa a mudança no desempenho na prova. Assim, se tivéssemos medido somente a ansiedade pré-prova e o desempenho na prova, poderíamos ter assumido que a alta ansiedade pré-prova causou péssimo desempenho na prova. Entretanto, está claro que péssimo desempenho na prova também pode ser explicado pela falta de revisão. Pode haver muitas variáveis adicionais que influenciam as variáveis correlacionadas e essas variáveis podem não ter sido medidas pelo pesquisador. Assim, poderia ser outra variável, não mensurada, que afeta o tempo de revisão e a ansiedade pré-prova. Field e Hole (2003) dão o exemplo da descoberta de Broca de um forte relacionamento entre gênero e tamanho do cérebro: o cérebro das mulheres são, em geral, menores do que os dos homens. Isso foi, naquela época, usado como evidência para argumentar que as mulheres eram intelectualmente inferiores aos homens. O problema óbvio é que esse relacionamento não leva em consideração o tamanho do corpo: pessoas com corpos maiores têm cérebros maiores independentemente da habilidade intelectual. Gould (1981) analisou novamente os dados e demonstrou que o forte relacionamento entre gênero e tamanho do cérebro desapareceu quando se levou em conta o tamanho do corpo!
- **Direção da causalidade:** Coeficientes de correlação nada dizem sobre qual variável causa a alteração na outra. Mesmo se pudéssemos ignorar o problema da terceira variável descrito acima, e pudéssemos assumir que as duas variáveis correlacionadas eram as únicas importantes, o coeficiente de correlação não indica em qual direção a causalidade opera. Assim, embora seja tentador concluir que a ansiedade pré-prova causa a mudança no desempenho na prova, não existe razão estatística de que o desempenho na prova não possa causar a mudança na ansiedade pré-prova. Embora essa última conclusão não faça sentido (porque a ansiedade foi medida antes do desempenho na prova), a correlação não nos diz que isso não é verdade.

4.5.3 Utilizando o R^2 para interpretação ①

Embora não possamos tirar conclusões diretas sobre causalidade, podemos levar o coeficiente de correlação um passo a frente elevando-o ao quadrado. O coeficiente de correlação ao quadrado (conhecido como o coeficiente de determinação, R^2) é uma medida da quantidade de variação em uma variável que é explicada pela outra. Por exemplo, podemos olhar o relacionamento entre ansiedade pré-prova e desempenho na prova. O desempenho na prova varia de pessoa para pessoa devido a diversos fatores (habilidades diferentes, diferentes níveis de preparação e assim por diante). Se adicionássemos todas as variabilidades (como quando calculamos a soma dos quadrados na Seção 1.4.1), teríamos uma estimativa de quanta variação existe no desempenho na prova. Podemos usar, então, o R^2 para nos dizer quanto dessa variabilidade é provocado pela ansiedade pré-prova. Essas duas variáveis tinham uma correlação de $-0{,}4410$ e, assim, o valor de R^2 será $(-0{,}4410)^2 = 0{,}194$. Esse valor nos diz quanto da variação no desempenho na prova pode ser explicado pela ansiedade pré-prova. Se convertermos esse valor em percentagem, podemos dizer que

Material com direitos autorais

ansiedade pré-prova tem 19,4% de variabilidade no desempenho na prova. Assim, embora a ansiedade pré-prova estivesse altamente correlacionada com o desempenho na prova, ela contribui somente com 19,4% da variação nos escores da prova. Colocando esse valor em perspectiva, isso nos leva a 80,6% da variabilidade ainda a ser contabilizada por outras variáveis. Devo ressaltar que embora R^2 seja uma medida útil da importância de um efeito, ele não pode ser usado para inferir relacionamentos causais. Embora geralmente falemos em termos de "a variância em *y* ser responsável por *x*" ou mesmo a variação em uma variável explicada pela outra, isso ainda nada diz sobre qual o caminho da causalidade. Assim, mesmo que a ansiedade pré-prova seja responsável por 19,4% da variação em escores de exames, ela não necessariamente causa a variação.

4.5.4 Coeficiente de correlação de Spearman ①

O coeficiente de correlação de Spearman, r_s, é uma estatística não-paramétrica e, assim, pode ser usada quando os dados violarem suposições paramétricas, tais como dados não-normais (veja o Capítulo 3). O teste de Spearman trabalha classificando os dados em primeiro lugar e então aplicando a equação de Pearson (equação (4.3)) aos dados ordenados. Como professor de estatística, estou sempre interessado em fatores que determinam se um estudante irá bem em um curso de estatística. Um fator consideravelmente importante é seu conhecimento prévio em matemática (no mínimo a experiência prévia irá determinar se uma equação faz sentido!). Imagine que eu peguei 25 alunos e olhei suas notas no curso de estatística no final do seu primeiro ano na universidade. Na Grã Bretanha, o estudante pode obter um conceito de primeira classe, um conceito de segunda classe superior, um conceito de segunda classe inferior, um de terceira classe, passar ou reprovar. Esses conceitos são categorias, mas eles têm uma ordem (um conceito de segunda classe superior é melhor do que um de segunda inferior). Eu poderia, também, perguntar a esses alunos que conceito eles obtiveram nas provas de matemática do GCSE. Na Grã Bretanha, GCSEs são provas escolares feitas com a idade de 16 anos que levam os conceitos A, B, C, D, E ou F. Novamente, esses conceitos são categorias que têm uma ordem de importância (um A é melhor do que todos os outros conceitos). Quando você tem categorias como essas que podem ser ordenadas de uma maneira significativa, os dados são denominados de *ordinais*. Os dados não são intervalares porque um conceito de primeira classe cobre 30% do intervalo (70% – 100%) enquanto um de segunda superior atinge somente 10% do intervalo (60% – 70%). Quando os dados foram medidos somente no nível ordinal eles são chamados de não-paramétricos e a correlação de Pearson não é apropriada. Portanto, o coeficiente de correlação de Spearman é usado. Os dados para esse estudo estão no arquivo **grades.sav.** Eles estão organizados em duas colunas: uma rotulada de **stats** e uma rotulada de **gcse.** Cada uma das categorias descritas acima foi codificada com um valor numérico. Em ambos os casos, o conceito mais alto (primeira classe ou conceito A) foi codificado com o valor de 1, com as categorias subsequentes rotuladas 2, 3 e assim por diante. Note que para cada valor numérico providenciei um rótulo (como fizemos para codificar variáveis).

O procedimento para determinar a correlação de Spearman (Figura 4.16) é o mesmo para a correlação de Pearson exceto que na caixa de diálogo das correlações bivariadas (Figura 4.15), precisamos selecionar ☑ Spearman e não a opção da correlação de Pearson. Nesse estágio, você deve também especificar se quer um teste uni ou bilateral. Para o exemplo acima, previ que as melhores notas no GCSE de matemática seriam correlacionadas com

Material com direitos autorais

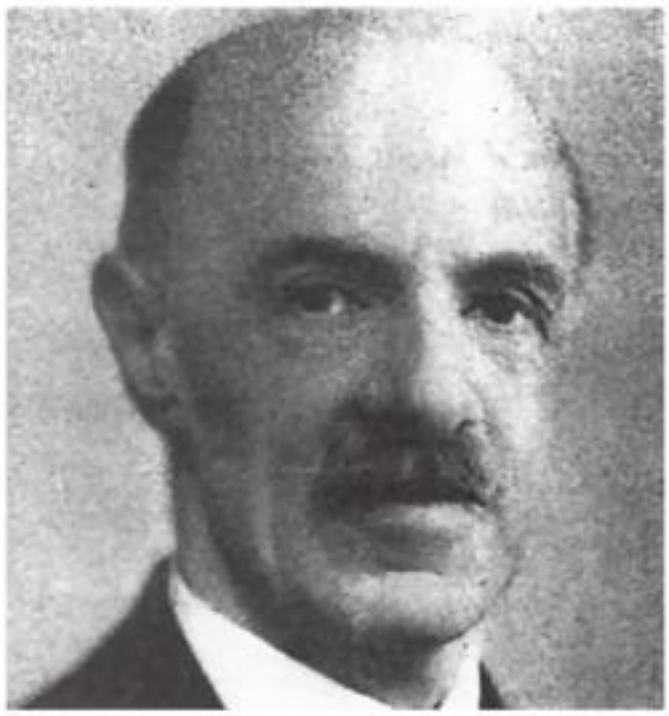


Figura 4.16 Charles Spearman, classificando furiosamente.

melhores notas no meu curso de estatística. A hipótese é direcional e um teste unilateral deve ser selecionado.

A Saída 4.3 do SPSS mostra a correlação de Spearman nas variáveis **stasts** e **gcse**. Essa saída é muito similar àquela da correlação de Pearson: uma matriz é mostrada dando o coeficiente de correlação entre duas variáveis (0,455), embaixo está o valor de significância desse coeficiente (0,011) e finalmente o tamanho da amostra (25). O valor da significância para esse coeficiente de correlação é menor do que 0,05; portanto, pode ser concluído que existe um relacionamento significativo entre os conceitos dos estudantes no GCSE de matemática e o seu conceito no curso de estatística. A própria correlação é positiva: assim, podemos concluir que à medida que os conceitos no GCSE aumentam, existe uma melhoria correspondente nos conceitos de estatística. Dessa maneira, a hipótese foi sustentada. Finalmente, é bom notar que o valor de *N* corresponde ao número de observações que fizemos. Se ele não corresponder, então dados podem ter sido excluídos por alguma razão.

4.5.5 Tau de Kendall (não-paramétrico) ①

A tau de Kendall, τ, é outra correlação não-paramétrica e deve ser usada em vez do coeficiente de Spearman quando você tem um conjunto pequeno de dados com um grande número de postos empatados. Isso significa que se você ordenar todos os escores e muitos deles apresentarem o mesmo posto, o tau de Kendall deve ser usado. Embora a estatística de Spearman seja a mais popular dos dois coeficientes, acreditamos que a estatística de Kendall é realmente uma melhor estimativa da correlação na população (veja Howell, 1997, p. 293). Dessa maneira, podemos estabelecer generalizações mais precisas da estatística de

Saída 4.3

***Correlations* (Correlações)**

			Statistics Grade (Nota em Estatística)	*GCSE Maths Grade* (Nota GCSE em Matemática)
Spearman's rho (Ro de Spearman)	*Statistics Grade* (Nota em Estatística)	*Correlation Coefficient* (Coeficiente de Correlação)	1.000	0.455*
		Sig. (1 tailed) (Sig. unilateral)	.	0.011
		N	25	25
	GCSE Maths Grade (Nota GCSE em Matemática)	*Correlation Coefficient* (Coeficiente de Correlação)	0.455*	1.000
		Sig. (1 tailed) (Sig. unilateral)	0.011	.
		N	25	25

*Correlation is significant at the .05 level (1-tailed) (A correlação é significativa ao nível de 0,05 (unilateral))

Material com direitos autorais

recode (recodificar) que é a mesma da variável **gender** (gênero)**,** exceto que a codificação é inversa (1 = fêmea, 0 = macho). Se você repetir o cálculo da correlação de Pearson usando **recode** (recodificar) em vez de **gender** (gênero)**,** descobrirá que o coeficiente de correlação agora se torna −0,378. O sinal do coeficiente é completamente dependente do código da categoria que você determinar e, assim, temos de ignorar toda a informação sobre a direção do relacionamento. Entretanto, podemos ainda interpretar R^2 como anteriormente. Assim, nesse exemplo, $R^2 = 0{,}378^2 = 0{,}143$. Portanto, podemos concluir que gênero contribui por 14,3% da variação do tempo gasto fora de casa.

Imagine agora que queremos converter a correlação bisserial por ponto no coeficiente de correlação bisserial (r_b) (porque alguns dos gatos machos foram castrados e, assim, pode haver um continuum de masculinidade subjacente à variável gênero). Devemos usar a equação (4.4) na qual P_1 é a proporção dos casos que caíram na categoria 1 (o número de gatos machos) e P_2 é a proporção dos casos que caíram na categoria 2 (o número de gatos fêmeas). Nessa equação, y é a ordenada da distribuição normal no ponto onde existe P_1% da área de um lado e P_2% no outro (isso ficará mais claro com um exemplo):

$$r_b = \frac{r_{pb}\sqrt{(P_1P_2)}}{y} \qquad (4.4)$$

Para calcular P_1 e P_2 simplesmente usamos o menu **Analyze⇒Descriptive Statistics⇒Frequencies** (Analisar⇒Estatística descritiva⇒Freqüências) e selecionamos a variável **gender** (gênero)**.** Não há necessidade de clicar em mais nenhuma opção, pois os valores por omissão lhe darão tudo o que você precisa saber (isto é, a percentagem dos gatos que são machos ou fêmeas). Isso resulta em 53,3% da amostra são fêmeas (isso é P_2) enquanto os restantes 46,7% são machos (isso é P_1). Para calcular y, usamos esses valores e os valores da distribuição normal exibidos no Apêndice A.1. Encontre a ordenada onde a curva normal é dividida em 0,467 como a porção menor e 0,533 como a poção maior (na verdade, teremos de usar os valores mais próximos daqueles que são 0,4681 e 0,5319, respectivamente). O valor da ordenada é 0,3977. Se substituirmos esses valores na equação (4.4), obteremos 0,475 (veja abaixo), o que é um valor mais alto do que a correlação bisserial por ponto (0,378). Portanto, a escolha do coeficiente de correlação pode fazer uma diferença substancial no resultado. Você deve, portanto, ter cuidado ao decidir se sua variável dicotômica tem uma continuidade subjacente ou se ela é uma verdadeira variável discreta.

$$\begin{aligned} r_b &= \frac{r_{pb}\sqrt{(P_1P_2)}}{y} \\ &= \frac{(0{,}378)\sqrt{(0{,}533\times 0{,}467}}{0{,}3977} \\ &= 0{,}475 \end{aligned}$$

4.6 CORRELAÇÃO PARCIAL ②

4.6.1 A teoria por trás da correlação parte e parcial ②

Mencionei anteriormente que existe um tipo de correlação que nos permite ver o relacionamento entre duas variáveis quando o efeito de uma terceira variável é constante. Por exemplo, análises dos dados da ansiedade pré-prova (no arquivo **ExamAnxiety.sav**) mostraram que o desempenho na prova está negativamente relacionado com a ansiedade pré-prova, mas positivamente relacionado ao tempo de revisão e o próprio tempo de revisão estava negativamente relacionado com a ansiedade pré-prova. Esse cenário é complexo, mas dado que sabemos que o tempo de revisão está relacionado a ambos, ansiedade pré-prova e desempenho na prova, então, se queremos uma medida pura do relacionamento entre ansiedade pré-prova e desempenho na prova, devemos levar em

Material com direitos autorais

conta a influência do tempo de revisão. Usando os valores de R^2 para esses relacionamentos, sabemos que a ansiedade pré-prova contribui com 19,4% da variância do desempenho na prova, que o tempo de revisão contribui com 15,7% da variância no desempenho na prova e que o tempo de revisão contribui com 50,2% da variância na ansiedade pré-prova. Se o tempo de revisão contribui com metade da variância na ansiedade pré-prova, parece viável que pelo menos alguns dos 19,4% da variância no desempenho na prova debitada à ansiedade é a mesma variância que contribui no tempo de revisão. Como tal, alguma variância no desempenho da prova explicada pela ansiedade pré-prova não é *única* e pode ser atribuída ao tempo de revisão. **Uma correlação entre duas variáveis em que os efeitos de outras variáveis são constantes é conhecida como uma correlação parcial.**

A Figura 4.17 ilustra o princípio por trás da correlação parcial. Na parte 1 do diagrama há uma caixa denominada *Exam Performance* (Desempenho na prova) que representa a variação total nos escores da prova (esse valor seria a variância do desempenho na prova). Existe, também, uma caixa que representa uma variação na ansiedade pré-prova (novamente, essa é a variância daquela variável). Já sabemos que a ansiedade pré-prova e o desempenho na prova compartilham 19,4% da sua variação (esse valor é o coeficiente de correlação ao quadrado). Portanto, as variações dessas duas variáveis se sobrepõem (porque elas compartilham a variância) criando uma terceira caixa (aquela com linhas diagonais). A sobreposição dessas caixas representando desempenho na prova e ansiedade pré-prova é a variância comum. Da mesma forma, na parte 2 do diagrama a variância compartilhada entre o desempenho na prova e tempo de revisão está ilustrada. O tempo de revisão compartilha 15,7% da variação nos escores na prova. Essa variação compartilhada é representada pela área de sobreposição (preenchida com linhas diagonais). Sabemos que o tempo de revisão e a ansiedade pré-prova têm em comum 50% das suas variações, portanto, é muito provável que alguma variação no desempenho na prova compartilhada pela ansiedade pré-prova sejam as mesmas das variâncias compartilhadas pelo tempo de revisão.

A parte 3 do diagrama mostra a figura completa. A primeira coisa a ser notada é que os quadros (ou caixas) representando ansiedade pré-prova e o tempo de revisão têm uma grande sobreposição (eles dividem 50% da variação). Mais importante, quando olhamos quanto o tempo de revisão e a ansiedade contribuem para o desempenho na prova, vemos que existe uma porção do desempenho na prova que é compartilhada por ambos, ansiedade e tempo de revisão (a área pontilhada). Entretanto, ainda existem pequenas partes da variância no desempenho na prova que são únicas para as outras duas variáveis. Assim, embora na parte 1 a ansiedade pré-prova compartilha um pedaço grande da variação com o desempenho na prova, alguma dessa sobreposição é também compartilhada com o tempo de revisão. Se removermos o pedaço da variância que é também dividida pelo tempo de revisão, teremos uma medida do relacionamento único entre desempenho na prova e ansiedade pré-prova. **Nós usamos correlações parciais para descobrir o tamanho do pedaço único da variância. Portanto, podemos determinar uma correlação parcial entre ansiedade pré-prova e desempenho na prova enquanto "controlamos" o efeito do tempo de revisão.** Da mesma forma, podemos determinar uma correlação parcial entre o tempo de revisão e o desempenho na prova "controlando" os efeitos da ansiedade pré-prova.

4.6.2 Correlação parcial utilizando o SPSS ②

Para determinar uma correlação parcial com os dados do desempenho na prova, selecione a opção *Correlate* (Correlacionar) do menu *Analyze* (Analisar) e, então, selecione *Partial* (Parcial) (**Analyze⇒Correlate⇒Partial**) (Analisar⇒Correlacionar⇒Parcial…) e uma caixa de diálogo semelhante a da Figura 4.18 será ativada. Essa caixa de diálogo lista todas

Material com direitos autorais

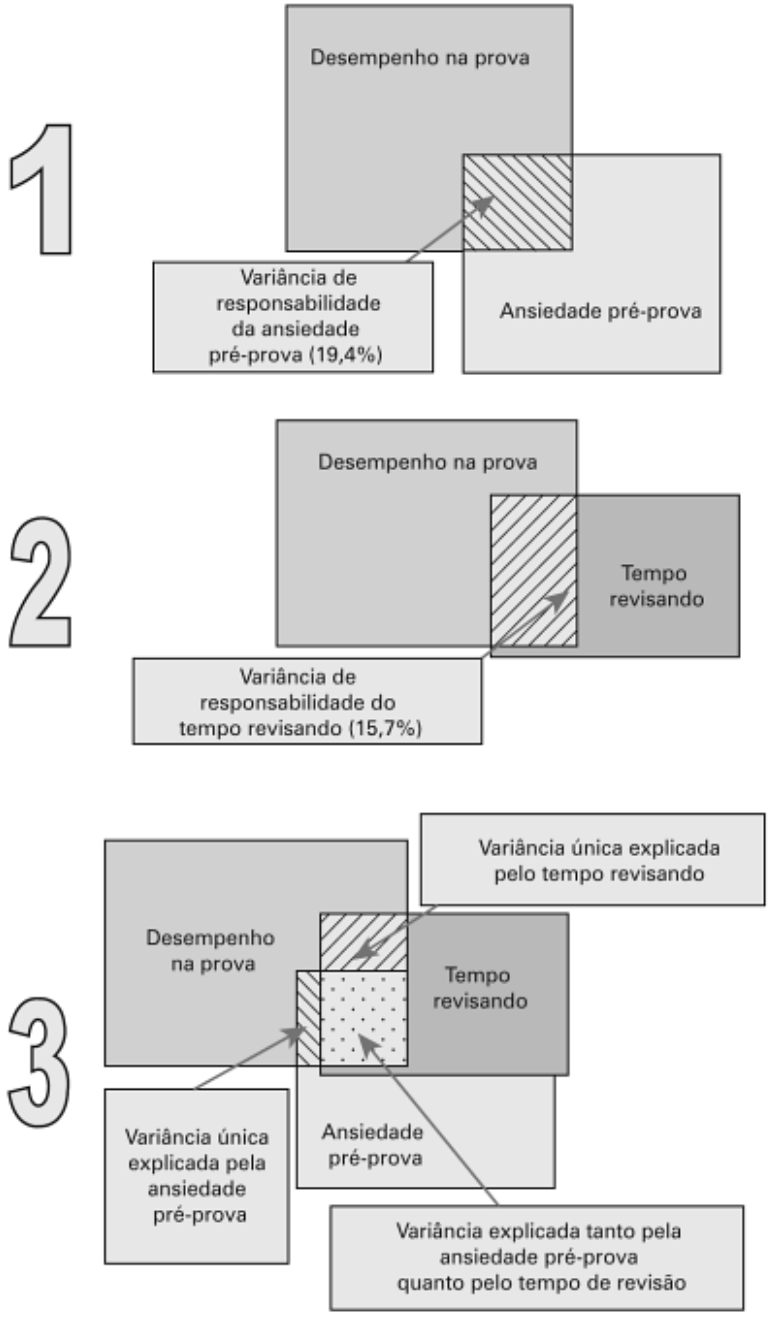


Figura 4.17 Diagrama mostrando os princípios da correlação parcial.

Material com direitos autorais

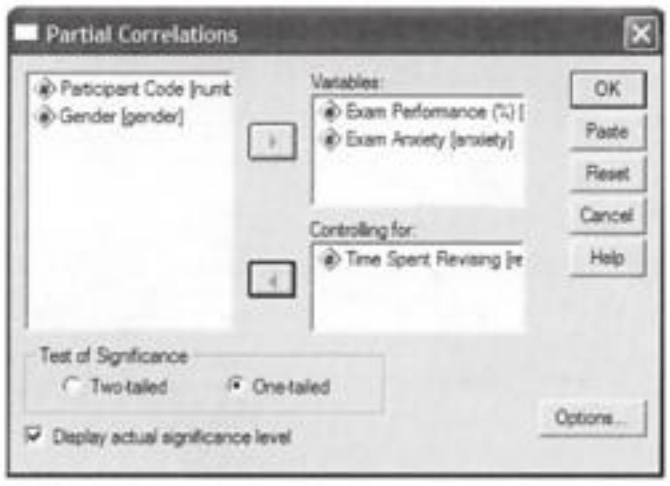


Figura 4.18 Caixa de diálogo principal para determinação de uma correlação parcial.

as variáveis do editor de dados no lado esquerdo deixando dois espaços vazios no lado direito. O primeiro espaço é para listar as variáveis que você quer correlacionar e o segundo é para indicar quais variáveis você quer controlar. No exemplo que descrevi, queremos determinar o efeito único da ansiedade pré-prova no desempenho na prova e, assim, queremos correlacionar as variáveis **exam** (prova) e **anxiety** (ansiedade), enquanto controlamos o tempo de revisão. A Figura 4.18 mostra a caixa de diálogos completa. Se você clicar em Options... (Opções) outra caixa de diálogo aparece como mostra a Figura 4.19.

Essas opções são similares àquelas da correlação bivariada exceto que você pode escolher mostrar correlações de ordem zero. Correlações de ordem zero são os coeficientes de correlação bivariada sem controle de quaisquer outras variáveis (isto é, são o coeficiente de correlação de Pearson). Assim, em nosso exemplo,

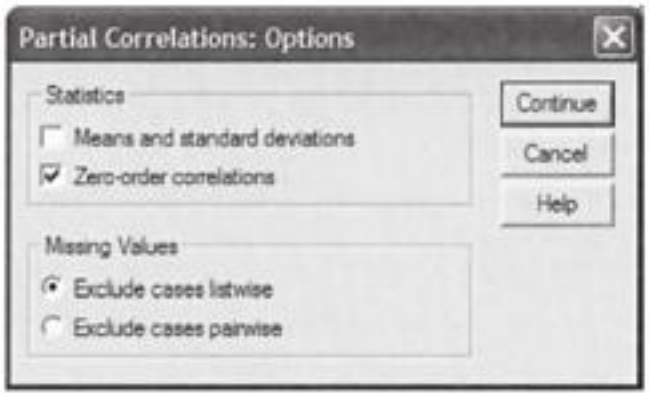


Figura 4.19 Opções para correlação parcial.

se selecionarmos as correlações de ordem zero, o SPSS irá produzir uma matriz de correlações entre **anxiety, exam,** e **revise** (ansiedade, prova e revisão). Se você não determinou as correlações bivariadas antes das correlações parciais, essa é uma maneira útil de comparar correlações que não foram controladas com as que foram. Essa comparação dá a você uma percepção da contribuição das diferentes variáveis. Marque a caixa das correlações de ordem zero, mas deixe o restante das opções como estão.

A saída 4.6 do SPSS mostra os resultados para a correlação parcial da ansiedade pré-prova e o desempenho na prova controlada pela variável tempo de revisão. A primeira coisa a notar é a matriz das correlações de ordem zero que foi obtida porque utilizamos a caixa de diálogos *options* (opções). As correlações mostradas aqui são idênticas àquelas obtidas com o procedimento da correlação de Pearson (compare essa matriz com a da saída 4.2 do SPSS). Abaixo das correlações de ordem zero está a matriz de correlações para as variáveis **anxiety** e **exam** (ansiedade e prova), mas controlada pela variável revisão. **Por ora controlamos uma única variável, assim, obtivemos o que é conhecido como uma correlação parcial de primeira ordem.** É possível controlar os efeitos de duas variáveis ao mesmo tempo (uma correlação parcial de segunda ordem) ou controlar três variáveis (uma correlação parcial de terceira ordem) e assim por diante. Em primeiro lugar, note que a correlação parcial entre desempenho na prova e ansiedade pré-prova é de $-0{,}2467$, que é consideravelmente menor do que a correlação quando o tempo de revisão não é controlado por ($r = -0{,}4410$). Na verdade, o coeficiente de correlação é aproximadamente menor do que ele era antes. Embora essa correlação seja estatisticamente significativa (seu valor p está abaixo de 0,05) o relacionamento diminuiu. Em termos de variância, o valor de R^2 para a correlação parcial é de 0,06, o que significa que a ansiedade pré-prova pode agora contribuir com somente 6% da variância no desempenho na prova. Quando os efeitos do tempo de revisão não eram controlados, a ansiedade pré-prova compartilhava 19,4% da

Material com direitos autorais

Saída 4.6 Saída para uma correlação parcial

```
PARTIAL CORRELATION COEFFICIENTS (COEFICIENTES DE CORRELAÇÃO PARCIAIS)

Zero Order Partials (Parciais de ordem zero)
                        EXAME       ANXIETY      REVISE
                       (Prova)   (Ansiedade)  (Revisar)
EXAME (Prova)           1.0000      -.4410       .3967
                       (     0)    (   101)    (   101)
                        P = .      P = 000      P = 000

ANXIETY (Ansiedade)     -.4410      1.0000       -.7092
                       (   101)    (     0)     (   101)
                      P = .000      P = .      P = .000

REVISE (Revisar)         .3967      -.7092       1.0000
                       (   101)    (   101)     (     0)
                      P = .000    P = .000      P = .

(Coeficientes / D.F.) / 1-tailed significance)
(Coeficientes / G.L. / Significância unilateral)

PARTIAL CORRELATION COEFFICIENTS (COEFICIENTES DE CORRELAÇÃO PARCIAIS)
Controling for .. (Controle por ..) REVISE(REVISAR)
                        EXAME       ANXIETY
                       (Prova)   (Ansiedade)
EXAME (Prova)           1.0000      -.2467
                       (     0)    (    100)
                        P = .      P = .006

ANXIETY (Ansiedade)     -.2467      1.0000
                       (    100)    (     0)
                      P = .006      P = .

(Coeficientes / D.F.) / 1-tailed significance)
(Coeficientes / G.L. / Significância unilateral)

". "is printed if a coefficient cannot be computed
(Se um coeficiente não puder ser calculado então ". "é impresso)
```

variância com os escores da prova e, assim, a inclusão do tempo de revisão diminuiu o montante da variação nos escores da prova compartilhados com a ansiedade. Dessa maneira, uma medida mais verdadeira do papel da ansiedade pré-prova foi obtida. A execução dessas análises tem mostrado que apenas a ansiedade pré-prova não explica muito da variância nos

Material com direitos autorais

4.9 TERMOS-CHAVE QUE DESCOBRIMOS

- Correlação bisserial
- Correlação bivariada
- Coeficiente de determinação
- Covariância
- Desvios dos produtos cruzados
- Tau de Kendall
- Padronização
- Correlação parcial
- Coeficiente de correlação de Pearson
- Correlação bisserial por ponto
- Diagrama de dispersão
- Correlação semiparcial
- Coeficiente de correlação de Spearman

4.10 TAREFAS DO ALEX ESPERTO

As respostas para essas tarefas podem ser encontradas no *site* www.artmed.com.br no arquivo **Answers-(Chapter 4).pdf**.

- Um estudante estava interessado em saber se havia um relacionamento positivo entre o tempo gasto escrevendo um artigo e a nota recebida. Ele pegou 45 dos seus amigos e contou quantas horas eles gastavam escrevendo seu artigo (**hours**) e a percentagem que eles tiveram no artigo (**essay).** Ele também traduziu essas notas na sua classificação de notas (**grade**): primeira, segunda superior, segunda inferior e terceira classe. Usando os dados no arquivo **EssayMarks.sav**, descubra qual foi o relacionamento entre o tempo gasto escrevendo o artigo e a nota final em termos de percentagem e classificação das notas (trace um diagrama de dispersão também!). ①
- Usando os dados do **ChickFick.sav** do Capítulo 3, existe um relacionamento entre gênero e excitação? Usando os mesmos dados, existe um relacionamento entre o filme assistido e excitação? ①

4.10 LEITURAS COMPLEMENTARES

HOWELL, D. C. *Statistical Methods forPsychology*. Belmont (CA): Duxbury, 2002 (5ª edição). O Capítulo 9 fornece uma cobertura mais detalhada sobre correlação e regressão do que Wright, mas a leitura não é tão fácil. O Capítulo 10 é ótimo para correlação bisserial e bisserial por ponto.

WRIGHT, D. B. *First Steps in Statistics*. Londres: Sage, 2002. Esse livro apresenta uma introdução muito clara aos conceitos de correlação e regressão (Capítulo 8).

Material com direitos autorais

5
REGRESSÃO

5.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ①

No último capítulo vimos como medir (e traçar) relacionamentos entre duas variáveis. Podemos avançar um passo nesse processo e aprender como prever uma das variáveis em função da outra. Um exemplo simples é o de tentar prever os níveis de estresse a partir do tempo que falta até a ocorrência de uma palestra. Você esperaria que esse fosse um relacionamento negativo (quanto menor o tempo até a palestra, maior a ansiedade). Podemos estender esse relacionamento básico para responder a questões como "se faltam 10 minutos até que alguém precise falar a um público, quão ansioso ele estará?". Essa é a essência da análise de regressão: é uma forma de prever algum tipo de saída (resultado) a partir de uma ou mais variáveis previsoras. Este capítulo aborda a análise de regressão com profundidade. Começo apresentando alguns princípios básicos utilizando o exemplo de prever um resultado a partir de um único previsor (**regressão simples**). Em seguida, explicarei o que é um modelo, como estimá-lo, como podemos determinar o quão bem ele se ajusta aos dados, como tudo isso é feito no SPSS e como as saídas do SPSS são interpretadas. Depois, analiso o caso mais complexo de quando um resultado é previsto a partir de diversas variáveis previsoras (**regressão múltipla**). Embora eu não aborde como o modelo é estimado, irei descrevê-lo e especificar vários métodos que selecionaram as variáveis que farão parte do modelo. Veremos, ainda, algumas coisas que podem afetar a precisão do modelo e como tratá-las antes de realizar a análise no SPSS, interpretar a saída e descobrir quão bem o modelo se ajusta aos dados e como ele pode ser generalizado. Se você ainda está acordado, o capítulo termina verificando como podemos incorporar variáveis categóricas na análise de regressão utilizando algo chamado de variáveis auxiliares (*dummy*). Ufa!

5.2 UMA INTRODUÇÃO À REGRESSÃO ①

A correlação pode ser uma ferramenta de pesquisa bastante útil, mas ela nada nos informa sobre o poder preditivo das variáveis. Na análise de regressão, ajustamos um modelo preditivo aos nossos dados e então usamos esse modelo para prever valores da variável dependente (VD) a partir de uma ou mais variáveis

Material com direitos autorais

independentes (VIs).[1] A regressão simples procura prever uma **variável de saída** a partir de um única variável previsora, enquanto que a regressão múltipla busca prever um resultado a partir de diversas **variáveis previsoras**. Essa é uma ferramenta bastante útil porque nos permite ir um passo além dos dados que de fato temos. Na Seção 1.4.11, apresentarei a ideia de que podemos prever qualquer valor utilizando a seguinte equação genérica:

$$\text{Saída}_i = (\text{Modelo}_i) + \text{Erro}_i \qquad (5.1)$$

Isso significa apenas que a saída ou resposta que estamos tentando prever para uma determinada pessoa pode ser prevista por qualquer modelo que ajustarmos aos dados mais algum tipo de erro. Na regressão, o modelo que ajustamos é linear ("modelo linear" significa "modelo baseado em uma linha reta") e você pode imaginar que está tentando resumir um conjunto de dados com uma linha reta (lembre a Figura 1.11). Assim, a palavra "modelo" na equação acima pode ser substituída por algo que defina a linha que ajustamos aos dados (veja a próxima seção).

Com qualquer conjunto de dados existem várias linhas que podem ser utilizadas para resumir a tendência geral e, desse modo, precisamos encontrar uma forma de decidir entre as muitas possibilidades. Para que seja possível tirar conclusões precisas, é necessário ajustar o *melhor* modelo que descreve os dados. Existem muitas maneiras de ajustar uma linha reta aos dados que foram coletados. A mais simples é utilizar seus olhos para imaginar uma linha que pareça resumir bem os dados. Entretanto, esse método é bastante subjetivo e não assegura que o modelo é o melhor que se poderia ter escolhido. Portanto, vamos utilizar técnicas matemáticas para determinar a linha que melhor descreve os dados observados. Esse método é denominado *método dos mínimos quadrados*.

5.2.1 Algumas informações importantes sobre retas ①

Mencionei acima que a palavra modelo na equação geral seria substituída por "alguma coisa que define a linha que se ajusta aos dados". De fato, qualquer linha reta pode ser definida por duas coisas: (1) a inclinação (ou gradiente) da linha (normalmente representado por b_1), e (2) o ponto em que a linha cruza o eixo vertical do gráfico (conhecido como *intercepto* da linha b_0). De fato, nosso modelo geral se torna a equação (5.2) na qual Y_i é a variável de saída que queremos prever e X_i é o escore do i-ésimo participante da variável previsora.[2] O gradiente da linha reta ajustada aos dados é b_1 e b_0 é o intercepto da linha. Os parâmetros b_1 e b_0 são conhecidos como *coeficientes de regressão* e aparecerão de vez em quando no livro, onde você irá vê-los normalmente representados como b (sem qualquer subscrito) ou b_i (que significa b associado à variável *i*). Existe um termo resíduo, ε_i, que representa a diferença entre o valor previsto pela linha para o participante i e o escore que o participante i realmente obteve. A equação é muitas vezes apresentada sem esse termo resíduo (assim, ignore-o se ele estiver lhe incomo-

[1] Infelizmente, você encontrará pessoas (e o SPSS também) se referindo às variáveis na regressão como dependentes e independentes (como em experimentos controlados). No entanto, a pesquisa correlacional por sua natureza raramente controla as variáveis independentes para medir o efeito na variável dependente. Em vez disso, as variáveis são mensuradas simultaneamente e sem um controle restrito. Assim, não é correto falar das variáveis de regressão dessa forma. Por esse motivo, denomino as "variáveis independentes" como *previsores* e a "variável dependente" como *saída* ou *resultado*.

[2] Às vezes, você irá ver essa equação escrita como:

$$Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$$

A única diferença é que nessa equação aparece β em vez de b e, de fato, as duas versões são a mesma coisa – elas apenas utilizam letras diferentes para representar os coeficientes. Eu prefiro a notação-β (e a utilizei na edição anterior); no entanto, quando o SPSS estima os coeficientes nessa equação, ele as rotula como b. Para ser consistente com a saída do SPSS que você verá, utilizarei b em vez de β.

Material com direitos autorais

dando); no entanto, vale a pena saber que esse termo significa que o nosso modelo não irá se ajustar perfeitamente aos dados coletados:

$$Y_i = (b_0 + b_1 X_i) + \varepsilon_i \qquad (5.2)$$

Uma linha particular tem um gradiente e um intercepto específicos. A Figura 5.1 mostra um conjunto de linhas que apresentam o mesmo intercepto, mas diferentes gradientes e outro conjunto que tem o mesmo gradiente (inclinação), mas diferentes interceptos. A Figura 5.1 também ilustra outro ponto importante: o de que o gradiente da linha nos informa algo sobre a natureza do relacionamento sendo descrito. No Capítulo 4, vimos que uma relação pode ser tanto positiva quanto negativa (e com isso não quero dizer que é a diferença entre se dar bem com a sua namorada ou discutir o tempo todo!). Uma linha que apresenta um gradiente positivo descreve um relacionamento positivo, ao passo que uma com um gradiente negativo descreve um relacionamento negativo. Assim, se você olhar para os gráficos na Figura 5.1 em que os gradientes diferem, mas os interceptos são os mesmos, notará que a linha mais grossa descreve um relacionamento positivo e a mais fina, um negativo.

Se é possível descrever uma linha conhecendo somente o gradiente (inclinação) e o intercepto dessa linha, então podemos utilizar esses valores para descrever o nosso modelo (porque no modelo de regressão linear utilizamos uma linha reta). Assim, é possível imaginar o modelo que ajustamos aos nossos dados na regressão linear como uma linha reta que pode ser descrita matematicamente pela equação (5.2). Com a regressão lutamos para encontrar a linha que melhor descreve os dados coletados, então estimamos o gradiente e o intercepto de tal linha. Tendo definido esses valores, podemos inserir diferentes valores na nossa variável previsora no modelo para estimar os valores da variável resposta (de saída).

5.2.2 O método dos mínimos quadrados ①

Já mencionei que o método dos mínimos quadrados é uma forma de encontrar a linha que melhor se ajusta aos dados (isto é, encontrar a linha que passe entre – ou o mais próximo possível – o maior número de pontos de dados). Essa "linha de melhor ajuste ou aderência" é encontrada ao verificarmos qual linha é a melhor de todas as possíveis que poderiam ser traçadas e resulta na menor diferença entre os pontos observados e a linha. A Figura 5.2 mostra que quando qualquer linha é ajustada a um conjunto de dados, existirá uma pequena diferença entre os valores previstos pela linha e os dados que foram efetivamente observados. Estamos interessados na diferença vertical entre a linha e os dados reais porque estamos utilizando a linha para prever valores de uma variável *Y* a partir de uma variável X. Embora alguns pontos de dados estejam exatamente sobre a linha, outros estão acima ou abaixo,

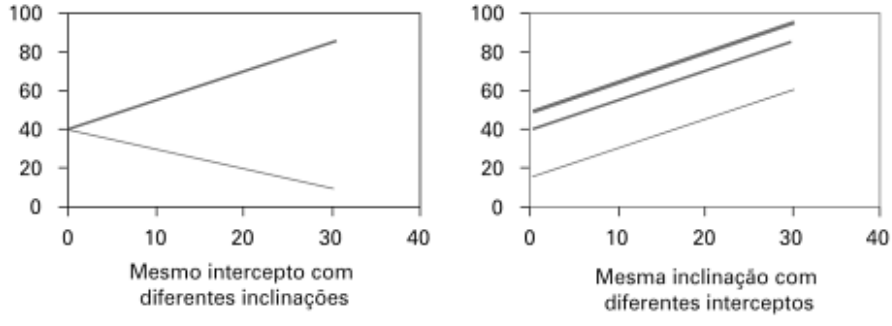


Figura 5.1 Linhas com os mesmos gradientes, mas com diferentes interceptos e linhas com o mesmo intercepto, mas com gradientes (inclinações) diferentes.

Material com direitos autorais

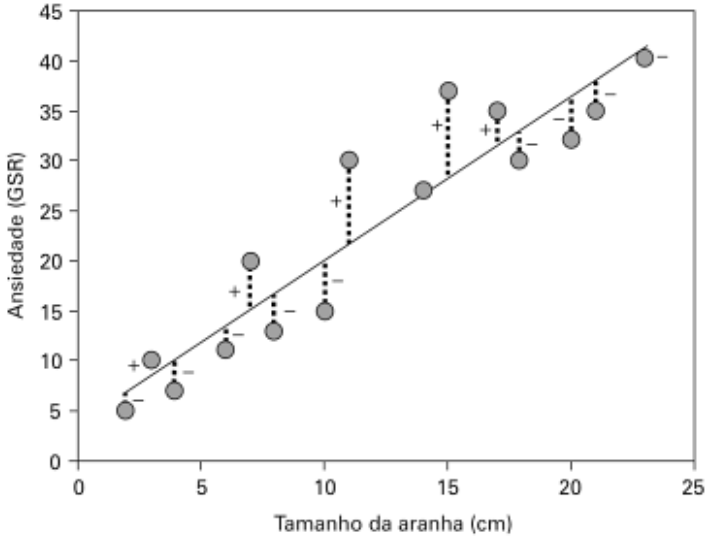


Figura 5.2 Esse gráfico mostra um diagrama de dispersão de alguns dados com uma linha de tendência entre eles. As linhas verticais (pontilhadas) representam as diferenças (resíduos) entre a tendência e os dados.

indicando que existe uma diferença entre o modelo que se ajusta aos dados e os dados coletados. Algumas dessas diferenças são positivas (os pontos estão acima da linha, indicando que o modelo está subestimando seus valores) e algumas são negativas (os pontos estão abaixo da linha, indicando que o modelo está superestimando seus valores). Essas diferenças são geralmente denominadas **resíduos**. Na discussão sobre variância na Seção 1.4.1, expliquei que se somarmos as diferenças positivas e negativas, elas tendem a se cancelar resultado numa soma zero. Para evitar esse problema, elevamos as diferenças ao quadrado antes de somá-las. As diferenças ao quadrado fornecem uma medida de quão bem uma determinada linha se ajusta aos dados: se as diferenças ao quadrado são grandes, a linha não é representativa dos dados; se as diferenças ao quadrado são pequenas, a linha é representativa. A soma das diferenças ao quadrado (ou SS abreviado) pode ser calculada para qualquer linha que é ajustada a dados; a qualidade de ajuste de cada linha aos dados pode ser avaliado pela soma das diferenças ao quadrado. O método dos mínimos quadrados seleciona a linha que apresenta a menor soma das diferenças ao quadrado (assim, ele escolhe a linha que melhor representa os dados observados). Uma forma de selecionar essa linha ótima seria ajustar cada linha possível ao conjunto de dados, calcular a soma das diferenças ao quadrado para cada uma e então escolher a linha que apresentasse o menor valor para essa soma. Você certamente gastaria um bom tempo fazendo isso! Felizmente, existe uma técnica matemática para encontrar máximos e mínimos e essa técnica (cálculo) é utilizada para encontrar a linha que minimiza a soma das diferenças ao quadrado. O resultado final é que os valores do intercepto e da inclinação da "linha de melhor aderência" podem ser estimados. Cientistas sociais geralmente se referem a essa linha de melhor ajuste como linha de regressão.

5.2.3 Avaliando o ajuste: soma dos quadrados, *R* e R^2 ①

Uma vez que encontramos a linha de melhor ajuste, é importante que avaliemos o quão bem ela adere aos dados (ou seja, a **qualidade do ajuste** do modelo). Fazemos isso porque

Material com direitos autorais

Dessa forma, ele representa o percentual de variação nos valores previstos que podem ser explicados pelo modelo:

$$R^2 = \frac{SS_M}{SS_T} \quad (5.4)$$

Não por coincidência, esse valor é o mesmo que encontramos para R^2 no Capítulo 4 (Seção 4.5.3) e você notará que ele é interpretado da mesma forma. Portanto, na regressão simples podemos tomar a raiz quadrada do seu valor para obter o coeficiente de correlação de Pearson. Como tal, o coeficiente de correlação nos fornece uma boa estimativa da aderência do modelo de regressão e R^2 nos dá uma boa medida do valor desse relacionamento.

Uma segunda utilização das somas dos quadrados para avaliar o modelo é por meio do teste F. Abordaremos esse teste em profundidade no Capítulo 8, mas, brevemente, o teste F tem por base a razão de melhoria devida ao modelo (SS_M) e a diferença entre o modelo e os dados observados (SS_R). De fato, em vez de utilizar a soma dos quadrados, tomamos a média dessas somas (denominadas **quadrados médios** ou MS). Para obter a média da soma dos quadrados é necessário dividir pelos graus de liberdade (isso é semelhante a calcular a variância a partir da soma dos quadrados – veja a Seção 1.4.1). Para o SS_M, o grau de liberdade é simplesmente o número de variáveis no modelo, e para o SS_R, ele é o número de observações menos o número de parâmetros sendo estimados (isto é, o número de coeficientes beta incluindo a constante). O resultado é a média dos quadrados para o modelo (MS_M) e a média dos quadrados dos resíduos (MS_R). Nesse estágio não é essencial que você entenda como as médias dos quadrados são derivadas (isso será explicado no Capítulo 8). Entretanto, é importante que você entenda que a razão F (equação 5.5) é uma medida do quanto o modelo melhorou na previsão de valores comparado com o nível de não precisão do modelo:

$$F = \frac{MS_M}{MS_R} \quad (5.5)$$

Se o modelo é bom, então esperamos que a melhoria na previsão devida ao modelo seja grande (assim, MS_M será grande) e a diferença entre o modelo e os dados observados serão pequenas (assim, MS_R será pequeno). Resumindo, um bom modelo deverá ter uma razão F grande (maior do que 1 pelo menos) porque o numerador da equação (5.5) será maior do que o denominador. A magnitude exata da razão F pode ser calculado utilizando valores críticos para os graus de liberdade correspondentes (como exemplificado no Apêndice A.3).

5.2.4 Interpretando previsões individuais ①

Vimos que o previsor em um modelo de regressão tem um coeficiente (b_1), que na regressão simples representa o gradiente da linha de regressão. O valor de b_1 representa a mudança na saída resultante da mudança de uma unidade no previsor. Se o modelo for inútil para previsão, então se o valor do previsor mudar, o que se poderá esperar da mudança no valor de saída? Bem, se o modelo é muito ruim, esperamos que essa alteração não ocorra, isto é, seja zero. Relembre da Figura 5.3 (veja o quadro representando SS_T) na qual vimos que utilizar a média não era a melhor forma de prever um resultado. De fato, a linha representando a média é horizontal, o que significa que à medida que a variável previsora muda, o valor previsto *não* se altera (porque para cada nível da variável previsora, prevemos que o valor de saída será o valor da média). O ponto importante aqui é que modelos ruins (como a média) terão um coeficiente de regressão igual a zero. Um coeficiente de regressão igual a zero significa: (1) a mudança de unidade na variável previsora não resultará em uma alteração no valor previsto (o valor previsto não muda de qualquer forma) e (2) o gradiente da linha de regressão é zero, assim, a linha de regressão será horizontal. A conseqüência lógica é que se uma variável prevê significativamente um valor de saída, então ela deve ter um valor

Material com direitos autorais

b_1 significativamente diferente de zero. Essa hipótese é testada utilizando o teste *t* (veja o Capítulo 7). A **estatística** *t* testa a hipótese nula de que o valor de b é zero: dessa forma, se for significativo aceitamos a hipótese que o valor de b é significativamente diferente de zero e que a variável previsora contribui significamente para a nossa habilidade de estimar o valor de saída.

Nosso problema em testar se o valor de b é diferente de zero é que ele depende da unidade de medida. Dessa forma, o teste *t* é calculado levando em conta o erro padrão. O erro padrão nos informa algo sobre como seria o comportamento dos valores b se fossem tomadas muitas amostras relativas a vendas de discos e investimento em publicidade e o valor de b fosse calculado para cada uma das amostras. Poderíamos fazer um diagrama da distribuição de freqüências dos valores b dessas amostras para descobrir se eles seriam semelhantes ou bem diferentes (lembre a Seção 1.6). Podemos utilizar o desvio padrão dessa distribuição (conhecido como *erro padrão*) como uma medida da semelhança dos valores b entre as amostras. Se o erro padrão é pequeno, significa que muitas amostras provavelmente têm valores b similares ao que foi obtido na amostra coletada (porque a variação entre as amostras é pequena). O teste *t* nos informa se o valor b é diferente de zero relativo à variação dos valores b para amostras de mesmo tamanho. Quando o erro padrão é pequeno, mesmo pequenos desvios de zero podem refletir uma diferença significativa porque b é representativo da maioria das possíveis amostras.

A equação (5.6) mostra como o teste *t* é calculado e você irá encontrar uma versão geral dessa equação no Capítulo 7 (equação (7.1)). O $b_{esperado}$ é simplesmente o valor de b que esperaríamos obter se a hipótese nula fosse verdadeira. Mencionei antes que a hipótese nula é de que b é zero, assim, esse valor pode ser substituído por zero.* A equação torna-se o valor observado de b dividido pelo erro padrão com qual ele está associado:[3]

$$t = \frac{(b_{observado} - b_{esperado})}{SE_b}$$

$$= \frac{b_{observado}}{SE_b}$$

Os valores de *t* seguem uma distribuição especial que diferem de acordo com um parâmetro denominado grau de liberdade. Na regressão, o grau de liberdade é simplesmente $N - p - 1$, onde *N* é o tamanho da amostra e p é o número de previsores. Na regressão simples, quando existe apenas um previsor, o grau de liberdade passa a ser $N - 2$. Tendo determinado que a distribuição *t* precisa ser utilizada, o valor observado de *t* pode ser comparado com valores que se esperaria obter apenas por acaso: se *t* é muito grande, é improvável que ele tenha ocorrido por acaso (esses valores podem ser encontrados no Apêndice A.2). O SPSS fornece a probabilidade exata de que o valor de *t* tenha ocorrido por acaso se o valor de b fosse de fato zero. Como regra geral, se esse valor (significância) for menor do que 0,05 ou 5%, os cientistas sociais concordam que b é significativamente diferente de zero; ou seja, o previsor contribui de forma significativa para o valor da saída (veja o Capítulo 1).

* N. de T.: Em geral, a hipótese nula é de que o coeficiente angular é zero, mas não necessariamente.

[3] Para comprovar isso você pode utilizar os valores da saída 5.3 do SPSS abaixo a fim de calcular *t* para a constante (intercepto) ($t = 134{,}140/7{,}537 = 17{,}79$) e investimento em publicidade (gradiente). Para o investimento em publicidade, você irá obter um valor de *t* diferente se utilizar os valores como os apresentados na saída do SPSS. Isso ocorre porque o SPPS arredonda os valores para 3 casas decimais nas tabelas, mas calcula o valor *t* sem arredondar esses valores (normalmente isso não acarreta diferenças substanciais, mas nesse caso sim!). Para obter os valores não-arredondados, clique duas vezes na tabela da saída do SPSS e então dê um duplo clique no valor que você quer ver sem arredondamento. Você deve encontrar que $t = 0{,}096124/0{,}009632 = 9{,}979$.

Material com direitos autorais

5.3 EXECUTANDO REGRESSÃO SIMPLES NO SPSS ①

Até agora, vimos um pouco da base teórica da regressão, embora restrita a casos em que existe apenas um previsor. Para auxiliar a esclarecer o que já aprendemos, iremos apresentar um exemplo de regressão simples utilizando o SPSS. Anteriormente, pedi que você imaginasse que eu trabalhava para uma gravadora e que meu chefe estava interessado em prever a venda de discos a partir da publicidade. Existem alguns dados para esse exemplo no arquivo **Record1.sav**. Esse arquivo de dados tem 200 linhas, cada uma representando um disco diferente. Existem também duas colunas, uma representando as vendas de cada disco na semana após o lançamento e outra representando a quantidade (em libras) gasta em promover o disco antes do seu lançamento. Esse é o formato para a entrada de dados de regressão: a variável de saída e a previsora devem estar em colunas diferentes e cada linha deve representar valores independentes dessas duas variáveis. O padrão desses dados é mostrado na Figura 5.4 e deve ficar claro que existe um relacionamento positivo: assim, quanto mais dinheiro for gasto na promoção de um disco, mais ele deve vender. É claro que alguns discos vendem bem a despeito do que for gasto em publicidade (canto superior esquerdo do diagrama), mas não existe um que venda muito mal quando muito dinheiro é gasto em publicidade (canto inferior direito do diagrama). O diagrama também mostra a linha de melhor aderência para esses dados: tendo em mente que a média é representada por uma linha horizontal em torno do valor de vendas de 200 000, a linha de regressão é visivelmente diferente.

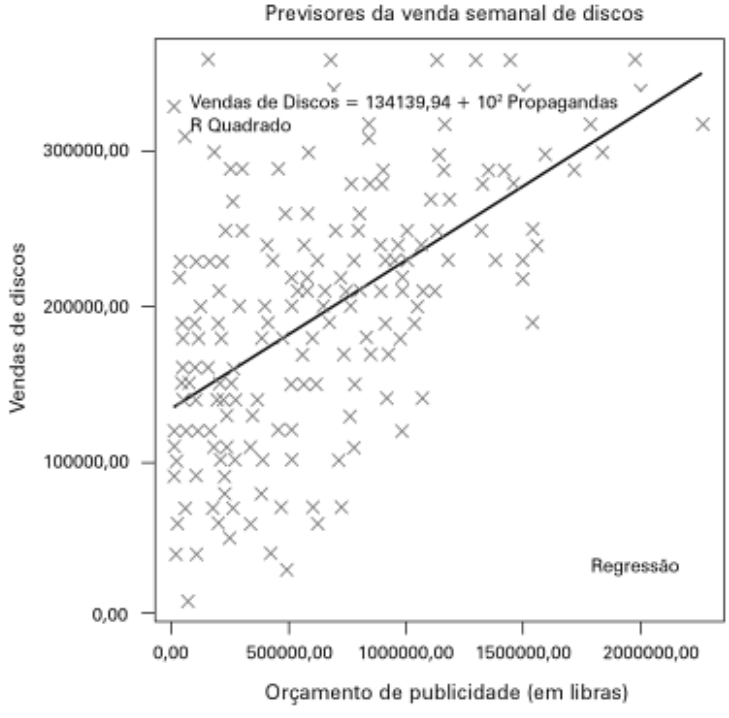


Figura 5.4 Diagrama de dispersão mostrando o relacionamento entre a venda de discos e a quantia gasta para promover o disco.

Material com direitos autorais

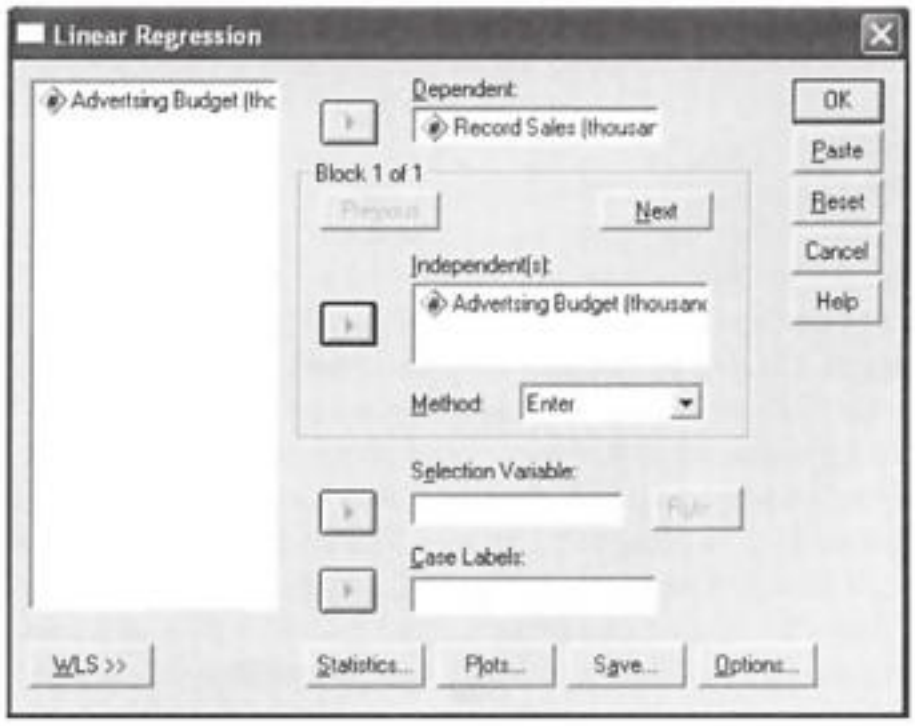


Figura 5.5 Caixa de diálogo principal para a regressão.

Para encontrar os parâmetros que descrevem a linha de regressão e para ver se essa linha é um modelo útil, precisamos executar uma análise de regressão. Para fazer a análise, você precisa acessar a caixa de diálogos principal utilizando o caminho **Analyze⇒Regression⇒Linear...** (Analisar⇒Regressão⇒Linear...). A Figura 5.5 mostra a caixa de diálogo resultante. Existe um espaço denominado *Dependent* (Dependente) no qual você deve colocar a variável de saída (nesse exemplo, **Sales** (Vendas)). Assim, selecione **sales** (vendas) da lista no lado esquerdo e transfira para o espaço clicando em ▸. Existe outro espaço rotulado de *Independent(s)* (Independentes) no qual cada variável previsora deve ser colocada. Na regressão simples, utilizamos uma única variável previsora, (nesse exemplo, **adverts** [publicidade]) então você deve selecionar **adverts** da lista clicando em ▸ de modo a transferi-la para a lista dos previsores. Existe uma variedade de opções disponíveis, mas elas serão exploradas dentro do contexto da regressão múltipla (veja a Seção 5.5). Por enquanto, clique em OK para executar a análise.

5.4 INTERPRETANDO A REGRESSÃO SIMPLES ①

5.4.1 Ajuste global do modelo ①

A primeira tabela fornecida pelo SPSS é um resumo do modelo. Essa tabela de resumo (Saída do SPSS 5.1) fornece os valores do R e do R^2 para o modelo que foi calculado. Para esses dados, R tem um valor de 0,578, e porque existe apenas um previsor, esse valor representa a correlação simples entre o valor investido em publicidade e a venda de discos (você pode confirmar isso executando uma correlação utilizando o que foi visto no Capítulo 4). O valor de R^2 é 0,335, o que nos informa que o gasto em publicidade pode ex-

Saída 5.1 do SPSS

***Model Summary* (Resumo do modelo)**

Model (Modelo)	*R*	R^2	*Adjusted R Square* (R quadrado ajustado)	*Standard Error of the Estimate* (Erro padrão da estimativa)
1	0.578[a]	0.335	0.331	65.9914

a. (Previsores: (constante), Orçamento de publicidade (milhares de libras)

Material com direitos autorais

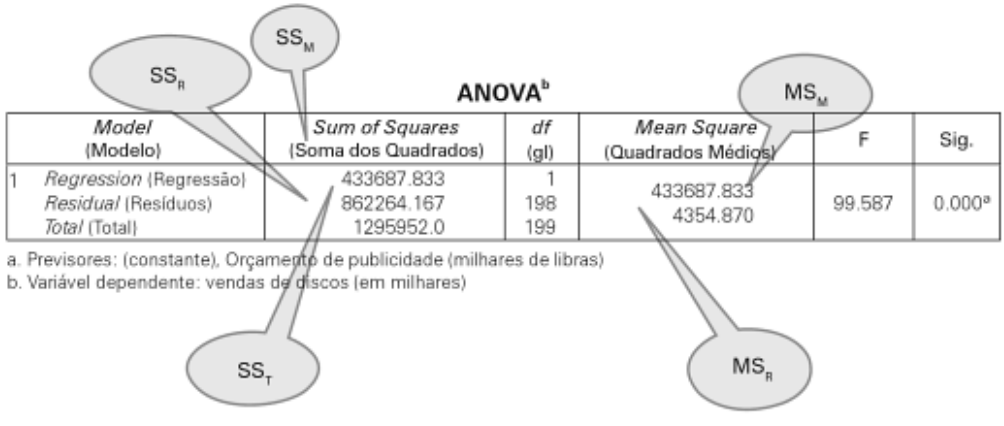
plicar 33,5% da variação das vendas de disco. Em outras palavras, se você está tentando explicar porque alguns discos vendem mais do que outros, podemos olhar a variação das vendas de discos diferentes. Devem existir muitos fatores que podem explicar essa variação, mas nosso modelo, que inclui somente a quantia gasta com publicidade, pode explicar 33% deles. Isso significa que 66% da variação das vendas não pode ser explicada apenas pela publicidade. Portanto, deve haver outras variáveis que também têm influência.

A próxima parte da saída relata uma análise de variância (ANOVA – veja o Capítulo 8). A tabela de resumo mostra as várias somas dos quadrados descritas na Figura 5.3 e os graus de liberdade associados com cada uma (Saída 5.2 do SPSS). A partir desses dois valores, a soma dos quadrados média pode ser calculada dividindo as somas dos quadrados pelo grau de liberdade associado. A parte mais importante da tabela é a razão F, que é calculada utilizando a equação (5.5) e a significância associada a esse valor F. Para esses dados, F é 99,59, que é significativo ao nível de $p < 0,001$ (porque o valor na coluna denominada *Sig.* é menor do que 0,001). Esse resultado nos informa que existe uma probabilidade menor do que 0,1% de que um valor F tão alto tenha ocorrido apenas por acaso. Desta forma, pode-se concluir que nosso modelo de regressão resulta em previsões melhores das vendas de disco do que se utilizássemos o valor médio das vendas. Resumindo, o modelo de regressão prevê as vendas de disco bastante bem.

5.4.2 Parâmetros do modelo ①

A ANOVA nos informa se o modelo, em geral, resulta em um grau de previsão significativamente bom dos valores da variável de saída. No entanto, a ANOVA não nos informa sobre a contribuição individual das variáveis no modelo (embora neste caso simples exista uma única variável no modelo e, assim, podemos inferir que esta variável é um bom previsor). A tabela na saída 5.3 do SPSS fornece detalhes dos parâmetros do modelo (os valores beta) e da significância desses valores. Vimos na equação (5.2) que b_0 é o intercepto y (ponto onde a linha corta o eixo *Y*) e esse é o valor B (na saída do SPSS) para a constante (intercepto). Assim, da tabela, podemos dizer que b_0 é 134,14 e isso significa que quando nenhum dinheiro for aplicado em publicidade (quando $X = 0$), o modelo prevê que 134 140 discos serão vendidos (lembre que a unidade de medida das vendas é em milhares de discos). Também podemos ler o valor de b_1 da tabela e esse valor representa a gradiente (inclinação)

Saída 5.2 do SPSS



ANOVA[b]

Model (Modelo)		Sum of Squares (Soma dos Quadrados)	df (gl)	Mean Square (Quadrados Médios)	F	Sig.
1	*Regression* (Regressão)	433687.833	1	433687.833	99.587	0.000[a]
	Residual (Resíduos)	862264.167	198	4354.870		
	Total (Total)	1295952.0	199			

a. Previsores: (constante), Orçamento de publicidade (milhares de libras)
b. Variável dependente: vendas de discos (em milhares)

Material com direitos autorais

anterior ao lançamento. O modelo existente que foi derivado utilizando o SPSS (veja a equação (5.7)) pode agora ser entendido para incluir essa nova variável (**airplay** – execuções):

$$\text{Venda de discos}_i = b_0 + b_1\text{Publicidade} + b_2\text{Execuções}_i + \varepsilon_i \qquad (5.10)$$

O novo modelo é baseado na equação (5.9) e inclui valores b para os dois previsores (e, é claro, a constante). Se calcularmos os valores b, podemos fazer previsões sobre a venda de discos não apenas utilizando a quantidade investida em publicidade, mas também em termos de execuções do disco na rádio. Existem somente dois previsores no modelo e, assim, ele pode ser apresentado graficamente em três dimensões (Figura 5.6).

A equação (5.9) descreve o trapézio escurecido no diagrama (conhecido como o *plano* de regressão), e os pontos destacados representam os pontos de dados observados. Como na regressão simples, o plano ajustado aos dados pretende ser o melhor possível para esses dados. No entanto, existem invariavelmente algumas diferenças entre o modelo e os dados da vida real (esse fato é evidente porque alguns dos pontos não residem exatamente sobre a área escurecida do gráfico). O valor b para a publicidade descreve a inclinação de cima para baixo do plano de regressão enquanto que o valor b para a execução (*airplay*) fornece a inclinação da esquerda para a direita do plano de regressão. Conhecer esses dois valores permite que coloquemos o plano de regressão no espaço.

É bastante simples visualizar o modelo de regressão com dois previsores utilizando um diagrama de dispersão tridimensional. Entretanto, a regressão múltipla pode ser utilizada com três, quatro ou mesmo mais variáveis.

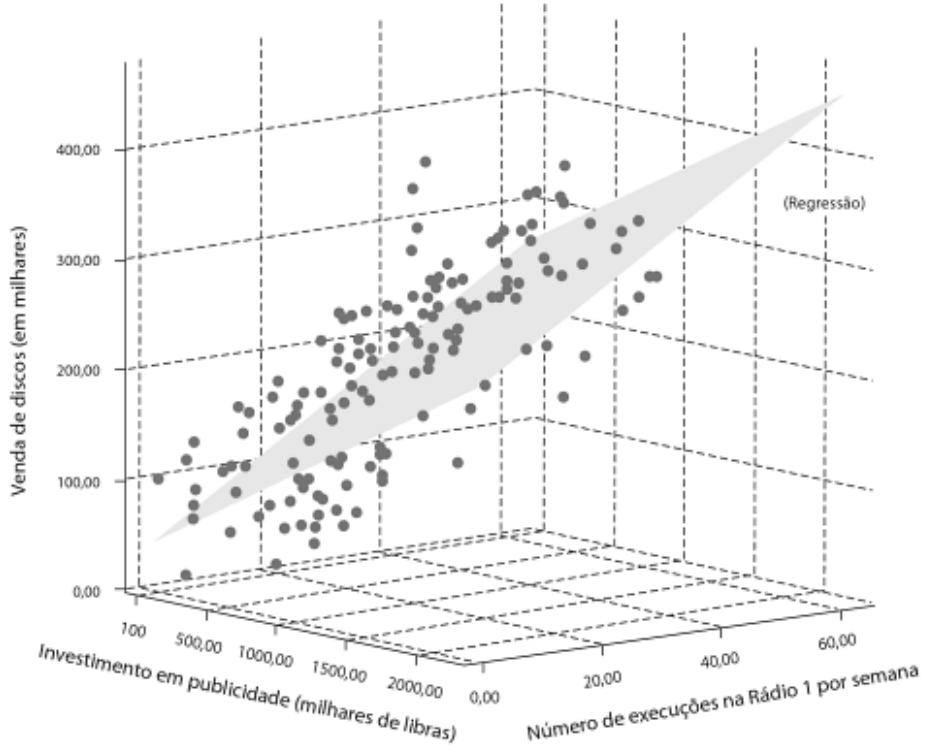


Figura 5.6 Diagrama de dispersão do relacionamento entre vendas de discos, investimento em publicidade e número de execuções do disco no rádio.

Material com direitos autorais

Embora você não possa imediatamente visualizar como tal modelo complexo se parece, ou o que os valores b representam, você deve ser capaz de aplicar os princípios desses modelos básicos a cenários mais complexos.

5.5.2 Soma dos quadrados, *R* e R^2 ②

Quando temos vários previsores, a partição da soma dos quadrados é a mesma que foi executada para uma única variável exceto que o modelo que estamos nos referindo toma a forma da equação (5.9) em vez de simplesmente ser uma linha reta bidimensional. Portanto, um SS_T pode ser calculado e representar a diferença entre os valores observados e a média dos valores da variável resultado. O SS_R ainda representa a diferença entre os valores previstos de *Y* pelo modelo e os valores observados de *Y*. Finalmente, SS_M pode ainda ser calculado e representa a diferença entre os valores de *Y* previstos pelo modelo e o valor da média. Embora o cálculo desses valores seja muito mais complicado que na regressão simples, conceitualmente esses valores são os mesmos.

Quando existem vários previsores, não faz sentido olhar para os coeficientes de correlação simples e, neste caso, o SPSS produz um coeficiente de correlação múltiplo (denominado ***R* Múltiplo**). O *R* Múltiplo é a correlação entre os valores observados de *Y* e os de *Y* previstos pelo modelo de regressão múltipla. Desta forma, valores grandes do *R* Múltiplo representam uma alta correlação entre os valores previstos e observados da variável de saída. Um *R* Múltiplo igual a 1 representa a situação na qual o modelo prediz com perfeição os valores observados, isto é, ele adere perfeitamente a todos os pontos. Como tal, o *R* Múltiplo é uma medida de quão bem o modelo prevê os dados observados. Segue que o R^2 resultante pode ser interpretado da mesma forma que na regressão simples: ele é a quantidade de variação na variável de saída que pode ser creditada ao modelo.

5.5.3 Métodos de regressão ②

Se você está interessado em construir um modelo complexo com vários previsores, como decidimos que previsores utilizar? Muito cuidado deve ser tomado ao selecionarmos previsores para um modelo porque os valores dos coeficientes de regressão dependem das variáveis utilizadas no modelo. Desta forma, os previsores incluídos e a forma que eles são colocados podem ter um grande impacto. Num mundo ideal, previsores deveriam ser selecionados baseados em pesquisas anteriores.[5] Se novos previsores estão sendo adicionados a modelos existentes, selecione aquelas novas variáveis com base na importância *teórica* substantiva de tais variáveis. Você não deve de forma alguma selecionar centenas de previsores ao acaso, juntá-los todos em uma análise de regressão e torcer pelo melhor. Além do problema da seleção de previsores, existem várias formas de entrada das variáveis no modelo. Quando os previsores são todos completamente não correlacionados, a ordem de entrada das variáveis tem pouco efeito nos parâmetros calculados; no entanto, nas pesquisas em ciências sociais raramente temos previsores não-correlacionados, assim, o método de seleção dos previsores é crucial.

5.5.3.1 Hierárquico (Entrada em blocos) ②

Na regressão hierárquica, os previsores são selecionados com base em trabalhos anteriores e o pesquisador decide em que ordem eles devem ser colocados no modelo. Como regra geral, previsores conhecidos (de outras análises) devem ser colocados no modelo primeiro pela sua ordem de importância em poder prever a variável de saída. Depois de entrar com todos os previsores conhecidos, o pesquisador pode adicionar os novos ao modelo. Novas variáveis podem ser colocadas de uma única vez, passo a passo ou hierarquica-

[5] Eu, um tanto cinicamente, apresento essa sugestão propondo que os previsores sejam escolhidos com base em pesquisas anteriores que tenham utilizado uma boa metodologia. Se você basear tais decisões nas análises de regressão, selecione previsores baseados somente em pesquisas passadas que utilizaram a regressão de forma apropriada e com modelos confiáveis e generalizáveis!

Material com direitos autorais

mente (de modo que o novo previsor que se julga mais importante entre primeiro).

5.5.3.2 Entrada forçada ②

Entrada forçada (ou *Enter*, como é conhecida no SPSS) é um método em que todos os previsores são forçados no modelo ao mesmo tempo. Da mesma forma que a hierárquica, este método baseia-se em boas razões teóricas para incluir os previsores escolhidos, mas diferentemente da hierárquica, o pesquisador não toma decisões sobre a ordem que as variáveis são acrescentadas.

5.5.3.3 Métodos por passos (Stepwise) ②

Na regressão passo a passo, as decisões sobre a ordem em que os previsores são acrescentados ao modelo é baseada puramente em critérios matemáticos. No método *forward* (para frente), um modelo inicial contendo somente a constante (b_0) é definido. O computador então procura pelo previsor (entre as variáveis possíveis) que melhor prevê a variável de saída – ele faz isso selecionando o previsor que apresenta o coeficiente de correlação simples mais alto com a variável de saída. Se esse previsor aumenta significativamente a habilidade do modelo prever a saída, ele é mantido no modelo e o computador procura por um segundo previsor. O critério para selecionar esse segundo previsor é procurar pela variável que apresenta a maior correlação semiparcial com a variável de saída. Deixe-me explicar isso de outra forma. Imagine que o primeiro previsor pode explicar 40% da variação da variável de saída; portanto, existe ainda 60% não explicada. O computador procura pelo previsor que pode explicar a mais alta taxa dos 60% restantes (assim, ele não está interessado nos 40% que já foram explicados). Em termos estatísticos, você pode pensar isso como uma correlação parcial na qual o computador correlaciona cada um dos previsores com a saída enquanto controla o efeito do primeiro previsor. A razão para chamá-la de correlação *semi*parcial é porque o efeito do primeiro previsor é parcializado apenas entre os previsores restantes e não são controlados na variável de saída. A correlação semiparcial fornece uma medida de quanto a "nova variância" na saída pode ser explicada por cada um dos previsores restantes (veja a Seção 4.6). O previsor que contribui mais para a nova variância é adicionado ao modelo, e se ele tiver uma contribuição significativa para o poder preditivo do modelo ele é mantido e outro previsor é procurado.

O método passo a passo (*stepwise*) do SPSS é o mesmo que o método *forward* (para frente), exceto que a cada vez que um previsor é adicionado a equação, um teste de remoção é feito sobre o previsor menos útil. Para tal a equação de regressão está constantemente sendo reacessada para ver se algum previsor redundante pode ser removido. O método para trás (*backward*) é o oposto do método para frente (*forward*), já que o computador inicia colocando todos os previsores no modelo e então calcula a contribuição de cada um verificando a significância do teste *t* de cada previsor. O valor da significância é comparado com um critério de remoção (que pode ser tanto o valor absoluto de uma estatística teste quanto o valor da probabilidade da ocorrência dessa estatística teste). Se o previsor satisfaz o critério de remoção (isto é, se ele não está fazendo uma contribuição estatisticamente significativa para o quão bem o modelo prevê a variável resultado) ele é removido do modelo e o modelo é reestimado com os previsores restantes. A contribuição dos demais previsores é então reavaliada.

Se você decidir utilizar o método passo a passo (*stepwise*), o método para trás (*backward*) é preferível ao método para frente (*forward*). Isso ocorre em virtude dos efeitos supressores, que acontecem quando um previsor tem um efeito significativo, mas somente quando outra variável é mantida constante. A seleção para frente (*forward*) tem uma probabilidade maior do que a para trás (*backward*) de excluir previsores envolvidos em efeitos supressores. Assim, com o método para frente (*forward*) há um risco maior de cometermos erro do Tipo II (isto é, eliminar um previsor que de fato contribui para o modelo).

Material com direitos autorais

5.5.3.4 Escolhendo um método ②

O SPSS permite que você opte por qualquer um desses métodos e é importante selecionar o mais apropriado. Os métodos para frente, para trás e passo a passo (*forward*, *backward* e *stepwise*) aparecem todos sob o título geral de métodos passo a passo (*stepwise methods*) porque em todos eles o computador escolhe as variáveis com base em um critério matemático. Muitos escritores argumentam que isso tira das mãos do pesquisador decisões metodológicas importantes. E mais, os modelos utilizados pelo computador freqüentemente tiram vantagem da variação da amostragem aleatória e, assim, decisões sobre que variáveis devem ser incluídas serão baseadas em pequenas diferenças nas correlações semiparciais. Entretanto, essas pequenas diferenças estatísticas podem contrastar dramaticamente com a importância teórica do previsor para o modelo. Por essa razão, é melhor evitar os métodos passo a passo (*stepwise*) exceto em uma análise exploratória do modelo (ver Wright, 1997, p. 181). Quando existe uma forte literatura teórica disponível, tome por base o que as pesquisas anteriores apontaram. Inclua qualquer variável significativa no modelo pela sua ordem de importância. Após essa análise inicial, repita a regressão, mas exclua qualquer variável que for estatisticamente redundante na primeira vez. Existem considerações importantes na decisão de quais previsores incluir. É importante não incluir muitos previsores. Como regra geral, quanto menos, melhor e certamente inclua somente previsores para os quais você tenha razões teóricas sólidas. Assim, seja seletivo e lembre que você deve ter um tamanho de amostra decente (alguns sugerem pelo menos 15 participantes por previsor – veja a Seção 5.6.2.3)

5.6 QUÃO ACURADO É O MEU MODELO DE REGRESSÃO? ②

Quando produzimos um modelo baseado em uma amostra dos dados existem duas questões importantes a serem feitas: (1) o modelo representa bem os dados, ou ele é influenciado por um pequeno número de casos e (2) o meu modelo pode ser generalizado para outras amostras? É fundamental fazer essas questões porque elas irão afetar como utilizaremos o modelo que está sendo construído. Essas questões são também, em certo sentido, hierárquicas, pois nós não queremos generalizar um modelo ruim. No entanto, é um erro pensar que se um modelo adere bem aos dados podemos tirar conclusões além da nossa amostra. **Generalização** é um passo adicional crítico; se acharmos que o nosso modelo não é generalizável, devemos restringir qualquer conclusão baseada no modelo da amostra utilizada. Primeiro vamos analisar como determinamos se um modelo é uma representação acurada dos dados reais e depois, na seção 5.6.2, verificaremos como é possível avaliar se um modelo pode ser usado para fazer inferências além da amostra dos dados que foi coletada.

5.6.1 Interpretando o modelo de regressão I: diagnósticos ②

Para responder a questão de quão bem o modelo adere aos dados observados ou se ele é influenciado por um pequeno número de casos, podemos buscar valores atípicos (*outliers*) ou casos influenciadores (a diferença é explicada no quadro 5.1). Veremos isso em etapas.

5.6.1.1 Valores atípicos (*outliers*) e resíduos ②

Um valor atípico é um caso que difere substancialmente da maioria dos dados (veja o Quadro 3.1). A Figura 5.7 mostra um exemplo de tal caso na regressão. Valores atípicos

Material com direitos autorais

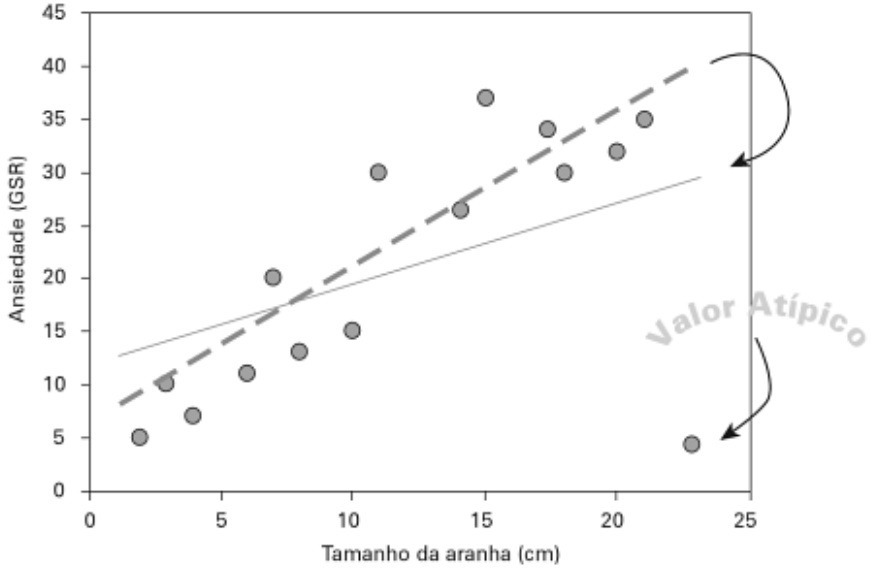


Figura 5.7 Gráfico demonstrando o feito de um valor atípico. A linha pontilhada representa a linha de regressão original para esses dados (veja a Figura 5.2),e a linha cheia representa a linha de regressão quando um valor atípico está presente.

podem introduzir tendenciosidade no modelo, pois eles irão afetar os valores dos coeficientes de regressão estimados. Por exemplo, a Figura 5.7 utiliza os mesmos dados da Figura 5.2, exceto que o escore de um participante foi alterado para ser um valor atípico (nesse caso, uma pessoa que ficou extremamente calma na presença de uma aranha muito grande). A mudança desse único valor teve um efeito dramático no modelo de regressão escolhido. Com o valor atípico presente, o modelo de regressão muda: a inclinação (gradiente) é reduzida (a linha se torna mais horizontal) e o intercepto aumenta (a linha cortará o eixo *Y* num ponto mais alto). Deve ficar claro a partir desse diagrama que é importante tentar detectar os valores atípicos para ver se o modelo é tendencioso.

Como você acha que poderia detectar um valor atípico? Sabemos que um valor atípico, pela sua natureza, é muito diferente de todos os demais escores. Assim, você acha que o modelo irá prever os escores dessas pessoas de forma acurada? Obviamente, a resposta é *não*: olhando para a Figura 5.7 é evidente que mesmo que o valor atípico tenha distorcido o modelo ele ainda prevê esse valor só que de uma forma muito ruim (a linha de regressão está bem longe dele). Portanto, se formos calcular as diferenças entre os valores dos dados coletados e os valores previstos pelo modelo, podemos detectar os valores atípicos buscando as diferenças grandes. Esse processo é o mesmo que procurar por casos em que o modelo não prevê de forma precisa. As diferenças entre os valores previstos pelo modelo e os valores observados na amostra são conhecidos como *resíduos*. Esses resíduos representam o erro que está presente no modelo. Se o modelo se ajusta bem aos dados da amostra, todos os resíduos devem ser pequenos (se o modelo aderir perfeitamente aos dados, todos os pontos estarão sobre a linha de regressão e todos os resíduos serão iguais a zero). Se o modelo não tiver uma boa aderência aos dados da amostra, os resíduos serão grandes. Além disso, se qualquer caso destacar-se por ter um grande resíduo, ele poderá ser atípico.

Os **resíduos** *normais* ou **não-padronizados** descritos acima são mensurados na mesma unidade da variável de saída e, portanto,

Material com direitos autorais

sora *X* prevê perfeitamente a variável de saída *Y* e não existirão erros. Se executarmos a análise novamente, mas desta vez incluindo o caso que não se ajusta (caso 30), o modelo resultante terá parâmetros diferentes. Alguns dados estão armazenados no arquivo **dfbeta.sav** que ilustra esta situação. Tente executar uma regressão simples primeiro com todos os casos incluídos e depois sem o caso 30. Os resultados estão resumidos na Tabela 5.1, que mostra (1) os parâmetros para o modelo de regressão quando o caso extremo está incluído ou excluído; (2) as equações de regressão resultantes; e (3) o valor de *Y* previsto para o escore do participante 30 na variável x (que é obtido substituindo o valor de *X* na equação de regressão pelo escore *X* do participante 30, que é 1).

Quando o caso 30 é excluído, esses dados apresentam um relacionamento negativo perfeito: o coeficiente para o previsor (b_1) é −1 (lembre que na regressão simples esse termo é o mesmo que o coeficiente de correlação de Pearson) e o coeficiente para a constante (o intercepto, b_0) é 31. Contudo, quando o caso 30 é incluído, os dois parâmetros são reduzidos[7] e a diferença entre os parâmetros também é mostrada. A diferença entre um parâmetro estimado utilizando todos os casos e estimado quando um caso é excluído é conhecida como **DFBeta** no SPSS. O DFBeta é calculado pra cada caso e para cada um dos parâmetros do modelo. Assim, no nosso exemplo, o DFBeta para a constante é −2 e o DFBeta para o previsor é 0,1. Olhando para esses valores do DFBeta, é possível identificar casos que tem uma grande influência nos parâmetros do modelo de regressão. Novamente, as unidades de medidas utilizadas afetarão esses valores, portanto, o SPSS produz um **DFBeta padronizado**. Esses valores padronizados são mais fáceis de usar porque pontos de corte universais podem ser aplicados. Nesse caso, valores absolutos acima de 1 indicam casos que substancialmente influenciam os parâmetros do modelo (embora Stevens, 1992, sugira olhar para os casos com valores absolutos maiores do que 2).

Uma estatística relacionada é a **DFFit**, que é a diferença entre o valor previsto para o caso quando o modelo é calculado incluindo o caso e quando o modelo é calculado excluindo o caso; neste exemplo, o valor é −1,09 (veja a Tabela 5.1). Se um caso não é influente, seu DFFit deve ser 0 – no entanto, é esperado que mesmo casos não-influentes tenham pequenos valores DFFit. Contudo, há o problema de que essa estatística depende da unidade de medida da variável de saída e, assim, um DFFit de 0,5 será muito pequeno se a saída variar de 1 a 100, mas muito grande se as saídas variarem de 0 a 1. Portanto, o SPSS também apresenta uma versão padronizada dos valores DFFit (**DFFit padronizados**). Uma medida final é a razão de covariância (CVR), que é uma medida de quanto um caso influencia a variância dos parâmetros de regressão. Uma descrição do cálculo dessa estatística deixaria muitos leitores tontos e confusos, portanto, é suficiente dizer que se essa razão está próxima de 1, o caso tem pouca influência nas variâncias dos parâmetros do modelo. Belsey, Kuh e Welsch (1980) recomendam o seguinte:

- Se $CVR_i > 1 + [(3(k+1)/n]$, então excluir o i-ésimo caso prejudicará a precisão de alguns dos parâmetros do modelo.

[7] O valor de b_i é reduzido porque os dados não apresentam mais um relação linear perfeita e existe agora uma variância que o modelo não pode explicar.

Tabela 5.1 A diferença entre os parâmetros do modelo de regressão quando um caso é excluído

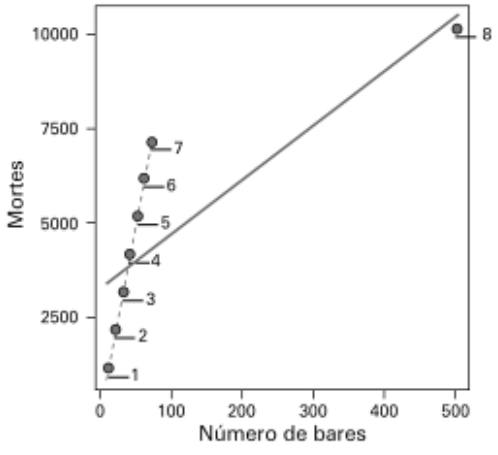
Parâmetros *b*	Caso 30 incluído	Caso 30 excluído	Diferença
Constante (Intercepto)	+29,00	31,00	−2,00
Previsor (Inclinação)	−0,90	−1,00	0,10
Modelo (Linha de regressão)	Y = (−0,9)X + 29	Y = (−1)X + 31	
Y previsto	28,10	30,00	−1,09

Material com direitos autorais

Quadro 5.1

A diferença entre resíduos e estatísticas de influência ③

Nesta seção, descrevemos duas maneiras de procurar casos que podem prejudicar o modelo: resíduos e as estatísticas de influência. Para ilustrar como essas medidas diferem, imagine que o prefeito de Londres do final do século passado estivesse interessado em como o ato de beber afetava a mortalidade. Londres é dividida em diferentes regiões denominadas municípios e, dessa forma, ele pode contar o número de bares e o número de mortes durante um período de tempo em oito dos seus municípios. Os dados estão em um arquivo denominado **pubs.sab**.



O diagrama de dispersão desses dados revela que sem o último caso, existe um relacionamento linear perfeito (a linha reta pontilhada). No entanto, a presença do último caso (caso 8) muda dramaticamente a linha de melhor ajuste (embora essa seja ainda um ajuste significativo dos dados – faça a análise de regressão e veja você mesmo).

É interessante olharmos para os resíduos e para as estatísticas de influência desses dados. O resíduo padronizado para o caso 8 é o segundo menor: esse valor atípico produz um resíduo bem pequeno (muitos dos dados típicos tem resíduos maiores) porque ele está bastante próximo da linha que foi ajustada aos dados. Como isso pode acontecer? Olhe para as estatísticas de influência e verás que elas são enormes para o caso 8, exercendo uma influência muito grande no modelo.

Como sempre, devemos fazer quando nos deparamos com estatísticas estranhas, é preciso perguntar o que está acontecendo no mundo real. O último ponto de dados representa a cidade de Londres, uma pequena área de apenas uma milha quadrada no centro de Londres onde poucas pessoas vivem, mas onde milhares de trabalhadores vem trabalhar e almoçam nos bares (*pubs*). Portanto, os bares não dependem de uma população que reside no local para fazer negócios e essa população não consumiu toda a sua cerveja! Dessa forma, havia um grande número de bares.

Isso ilustra que um caso que influencia muito pode produzir um pequeno resíduo – portanto, olhe ambos! (Agradeço a David Hitchin por esse exemplo, que ele obteve do Dr. Richard Roberts.)

***Case summaries*[a] (Resumo dos casos)**

		Standardized Residual (Resíduos padronizados)	*Centred Leverage Value* (Valor da Influência Centrado)	*Standardized DFFit* (DFFit padronizados)	*Standardized DFBETA Intercept* (Intercepto DFBETA padronizado)	*Standardized DFBETA PUBS* (DFBETA Bares padronizado)
1		–1.33839	0,04074	–0,74402	–0,74317	0,36886
2		–0,87895	0,03196	–0,40964	–0,40766	0,18484
3		–0,41950	0,02424	–0,17697	–0,17494	0,07132
4		0,3995	0,01759	0,01606	0,01572	–0,00564
5		0,49940	0,01200	0,20042	0,19337	–0,05933
6		0,95885	0,00748	0,40473	0,38333	–0,09618
7		1,41830	0,00402	0,68084	0,62996	–0,12023
8		–0,27966	0,86196	–460379232,7	92676016,019	–430238878,2
Total	N	8	8	8	8	8

a Limitado aos primeiros 100 casos.

Material com direitos autorais

- Se $CVR_i < 1 - [(3(k+1)/n]$, então excluir o i-ésimo caso irá melhorar a precisão de alguns dos parâmetros do modelo.

Nas duas expressões, k é o número de previsores do modelo, CVR_i é a razão da covariância para o i-ésimo participante e n é o tamanho da amostra.

5.6.1.3 Um comentário final sobre estatísticas diagnóstico ②

Existem muitas estatísticas diagnóstico que devem ser examinadas após a execução de uma análise de regressão e é difícil resumir essa riqueza de material em uma conclusão concisa. Contudo, eu gostaria de salientar um ponto discutido por Belseu e colaboradores (1980), que notou o perigo inerente nesses procedimentos. O problema é que os diagnósticos são recursos que permitem ver quão bom ou ruim é o seu modelo em termos da aderência aos dados da amostra. Eles são uma forma de avaliar o seu modelo. Eles não são, contudo, uma forma de justificar a remoção de dados que afetam algumas mudanças desejáveis nos parâmetros da regressão (por exemplo, eliminar um caso que muda um valor b não-significativo para um significativo). Stevens (1992), como sempre, oferece um excelente conselho:

> Se um ponto é um significativo valor atípico em Y, mas a sua distância de Cook é menor do que um, não existe motivo real para eliminar esse ponto, uma vez que ele não apresenta um grande efeito na análise de regressão. Contudo, existe ainda o interesse de estudar tal ponto para tentar entender por que ele não está aderindo ao modelo (p. 118).

5.6.2 Interpretando o modelo de regressão II: generalização ②

Quando uma análise de regressão é feita, uma equação correta para os valores da amostra observada pode ser produzida. Entretanto, nas ciências sociais normalmente estamos interessados em generalizar nossas descobertas para além da amostra observada. Assim, embora possa ser útil tirar conclusões sobre uma amostra particular de pessoas, geralmente é mais interessante se pudermos assumir que nossas conclusões são verdadeiras para uma população mais ampla. Para generalizar um modelo de regressão, devemos estar seguros de que as suposições foram satisfeitas, e para testar se o modelo de fato é generalizável, podemos fazer uma validação cruzada.

5.6.2.1 Avaliando hipóteses ②

Para tirar conclusões sobre uma população com base em um modelo de regressão realizado sobre uma amostra, várias suposições (hipóteses) devem ser verdadeiras (veja Berry, 1993).

- **Tipos de variáveis**: Todas as variáveis previsoras devem ser quantitativas ou categóricas (com duas categoriais), e a variável de saída deve ser quantitativa, contínua e não-limitada. Quantitativa significa que ela deve ser mensurada por intervalos e não-limitada quer dizer que não devem existir restrições na variabilidade da saída. Se a saída é uma medida que varia de 1 a 10 e os dados coletados variam entre 3 e 7, então esses dados são restritos.
- **Variância não-nula**: Os previsores devem ter alguma variação nos valores (isto é, eles não devem ter variância zero).
- **Multicolinearidade não deve ser perfeita**: Não deve existir relacionamento linear perfeito entre dois ou mais previsores. Assim, as variáveis previsoras não devem apresentar correlações muito altas (veja a Seção 5.6.2.4).
- **Previsores não-correlacionados com "variáveis externas"**: *Variáveis externas* são variáveis que não foram incluídas no modelo de regressão e que influenciam a variável de saída.[8] Essas variáveis podem

[8] Alguns autores se referem a essas variáveis externas como parte de um termo erro que inclui qualquer fator aleatório no modo pelo qual a saída varia. Contudo, para evitar confusão com os termos resíduos na equação de regressão, escolhi denominá-las "variáveis externas". Embora esse termo implicitamente anule qualquer fator aleatório, eu reconheço esses fatores aqui!

Material com direitos autorais

ser pensadas como semelhantes a uma "terceira variável" que foi discutida na correlação. Essa suposição significa que não devem existir variáveis externas que se correlacionam com qualquer uma das variáveis incluídas no modelo de regressão. Obviamente, se as variáveis externas se correlacionam com os previsores, as conclusões feitas com o modelo não serão confiáveis (porque outras variáveis também podem prever a saída de forma precisa).

- **Homocedasticidade**: A cada nível das variáveis previsoras, a variância do termo residual deve ser constante. Isso significa que os resíduos a cada nível dos previsores devem ter a mesma variância (homocedasticidade): quando as variâncias são desiguais diz-se que existe heterocedasticidade (veja a Seção 3.6 também).
- **Erros independentes**: Para quaisquer duas observações os termos resíduos devem ser não-correlacionados (ou independentes). Essa eventualidade é algumas vezes descrita como falta de **autocorrelação**. Essa suposição pode ser verificada com o **teste de Durbin-Watson**, que testa a correlação serial entre erros. Especificamente, ele testa se resíduos adjacentes são correlacionados. A estatística teste pode variar entre 0 e 4, com 2 significando que os resíduos não são correlacionados. Um valor maior do que 2 indica correlação negativa entre resíduos adjacentes e um valor abaixo de 2 indica uma correlação positiva. O tamanho da estatística de Durbin-Watson depende do número de previsores do modelo e do número de observações. Para ser acurado, você deve olhar os valores aceitáveis exatos no artigo original de Durbin e Watson (1951). Uma regra bastante conservadora, afirma que valores menores do que 1 ou maiores do que 3 são definitivamente motivos de preocupação; contudo, valores próximos de 2 também podem ser problemáticos dependendo da sua amostra e modelo.
- **Erros normalmente distribuídos**: Presume-se que os resíduos em um modelo são variáveis aleatórias, normalmente distribuídas com média zero. Essa hipótese significa que as diferenças entre o modelo e os dados observados são com mais frequência zero ou muito próximas a zero e que diferenças muito maiores do que zero acontecem apenas ocasionalmente. Algumas pessoas confundem essa hipótese com a ideia de que os previsores têm que ser normalmente distribuídos. Na verdade, os previsores não precisam ser normalmente distribuídos (veja a Seção 5.10).
- **Independência**: Assume-se que todos os valores da variável de saída são independentes (em outras palavras, cada valor da variável de saída provém de uma entidade separada).
- **Linearidade**: Os valores médios da variável de saída para cada incremento nos previsores devem estar sobre a linha. Em outras palavras, isso significa que é assumido que o relacionamento que estamos modelando é do tipo linear. Se modelarmos um relacionamento não-linear utilizando um modelo linear, isso obviamente limita a generalização do que encontrarmos.

Essa lista de hipóteses parece desestimulante e, de fato, muitos alunos de graduação (e alguns professores também) tendem a considerar as hipóteses como um assunto tedioso com o qual ninguém deveria ser preocupar. Quando menciono as suposições estatísticas para psicólogos, eles tendem a me lançar um olhar do tipo "você é realmente pedante" e então me ignoram. Contudo, existem boas razões para levar as suposições a sério. Imagine que eu vou à casa de um amigo e as luzes estão acesas e é óbvio que alguém está em casa. Eu toco a campainha da porta e ninguém atende. A partir dessa experiência, concluo que o meu amigo me odeia e que sou uma pessoa terrível e mal-amada. Até que ponto essa conclusão é verdadeira? Bem, existe uma realidade que estou tentando ana-

Material com direitos autorais

lisar (isto é, se o meu amigo gosta ou não de mim) e eu coletei dados sobre essa realidade (eu fui à casa dele, vi que ele está em casa, toquei a campainha e não obtive resposta). Imagine que na verdade o meu amigo gosta de mim (ele nunca soube escolher bem os seus amigos!); neste cenário, minha conclusão é falsa. Por que os meus dados me levaram a uma conclusão errada? A resposta é simples: eu assumi que a campainha da porta do meu amigo estava funcionando e sob essa suposição a conclusão que fiz a partir dos meus fatos foi precisa (meu amigo ouviu a campainha, mas escolheu me ignorar porque ele me odeia). Contudo, essa suposição não era verdadeira – sua campainha não estava funcionando e foi por isso que ele não respondeu – e, como conseqüência, a conclusão que fiz sobre a realidade foi completamente falsa.

Chega de campainhas, amigos e minha vida social: o ponto a ser lembrado é que, quando as hipóteses são ignoradas, nós paramos de tirar conclusões válidas sobre a realidade. Em termos de regressão, quando as suposições são consideradas, o modelo que obtemos de uma amostra pode ser aplicado de forma precisa para a população de interesse (os coeficientes da equação de regressão não são *tendenciosos*). Algumas pessoas assumem que isso significa que quando as hipóteses são satisfeitas, o modelo de regressão de uma amostra é sempre idêntico ao modelo que seria obtido se fôssemos capaz de testar toda a população. Infelizmente, essa crença não é verdadeira. O que um modelo não tendencioso nos diz é que, em *média*, o modelo de regressão obtido a partir de uma amostra é o mesmo que o modelo populacional. Entretanto, deve ficar claro que mesmo quando as suposições são satisfeitas, é possível que um modelo obtido a partir de uma amostra não seja igual ao modelo populacional – mas a probabilidade de serem idênticos aumenta.

5.6.2.2 Validação cruzada do modelo ③

Mesmo se não pudermos ter certeza de que o modelo derivado da nossa amostra representa de forma precisa toda a população, existem maneiras de determinar quão bem nosso modelo pode prever a saída em uma amostra diferente. Determinar a precisão de um modelo entre diferentes amostras é conhecido como **validação cruzada**. Se um modelo pode ser generalizado, ele deve ser capaz de prever de modo preciso a mesma variável de saída a partir do mesmo conjunto de previsores em um grupo de pessoas diferentes. Se o modelo é aplicado a uma amostra distinta e existe uma grande diferença na sua capacidade de previsão, então o modelo claramente não é generalizável. Como uma primeira regra prática, nós devemos coletar dados suficientes para obter um modelo de regressão confiável (veja a próxima seção). Uma vez determinado o modelo de regressão, existem dois métodos principais de validação cruzada:

- **R^2 ajustado**: No SPSS, não apenas os valores de R e R^2 são calculados, mas também um R ajustado. Esse valor ajustado significa a perda do poder de previsão ou *encolhimento*. O R^2 informa quanto da variância de Y pode ser creditada ao modelo de regressão amostral, e o valor ajustado nos informa quanta variância de Y pode ser creditada ao modelo se ele tiver sido derivado da população de onde a amostra foi retirada. O SPSS determina o valor R^2 ajustado utilizando a equação de Wherry. Contudo, essa equação tem sido criticada porque nada informa sobre quão bem o modelo de regressão preveria um conjunto de dados inteiramente diferente (quão bem o modelo prevê escores de diferentes amostras retiradas da mesma população?). Uma versão do R^2 que nos informa quão bem o modelo valida de forma cruzada utiliza a fórmula de Stein, mostrada na equação (5.11) (veja Stevens, 1992):

$$R_2 \text{ ajustado} = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right](1-R^2) \quad (5.11)$$

Material com direitos autorais

camente inevitável. Níveis baixos de colinearidade não representam uma ameaça para os modelos gerados pelo SPSS, mas à medida que a colinearidade aumenta, também aumentam os erros padrões dos coeficientes b, que por sua vez afetam a significância estatística desses coeficientes. Resumindo, altos níveis de colinearidade aumentam a probabilidade de que um bom previsor da variável de saída seja declarado não-significativo e excluído do modelo (Erro do Tipo II). Existem três outras razões por que a presença de multicolinearidade representa uma ameaça à validade de uma análise de regressão múltipla:

- **Ela limita o tamanho do R**: Lembre que R é uma medida de correlação múltipla entre os previsores e a saída e que R^2 indica a variância da saída que é de responsabilidade dos previsores. Imagine uma situação em que uma única variável prevê a variável de saída com um sucesso razoável (por exemplo, $R = 0{,}80$) e, depois, uma segunda variável previsora é adicionada ao modelo. Essa segunda variável deve ser responsável por boa parte da variância dos resultados (que é o motivo de ela ter sido acrescentada ao modelo), mas a variância que pode ser debitada a ela é a mesma da primeira variável. Em outras palavras, uma vez que a variância debitada à primeira variável tenha sido removida, o segundo previsor torna-se responsável por muito pouco da variância restante (a segunda variável apresenta uma variância própria muito pequena). Assim, a variável total dos resultados que pode ser debitada aos dois previsores é um pouco maior apenas do que quando um único previsor estava no modelo (assim, R pode aumentar de 0,80 para 0,82). Essa ideia está conectada à noção de correlação parcial que foi explicada no Capítulo 3. Se, contudo, os dois previsores são completamente não-correlacionados, o segundo previsor tem uma probabilidade maior de ser responsável por uma variância dos resultados do que quando existia um único previsor. Dessa forma, embora por si só o segundo previsor seja responsável apenas por uma pequena parte da variância dos resultados, a variância debitada aos dois é diferente do que quando tinha apenas um previsor (portanto, quando ambos estão no modelo, R aumenta substancialmente, por exemplo, para 0,95). Desse modo, previsores não-correlacionados são desejáveis.
- **Importância dos previsores**: A multicolinearidade entre previsores dificulta a avaliação da importância individual de um previsor. Se os previsores são altamente correlacionados e cada um é responsável por uma variância similar dos resultados, como podemos saber qual entre duas variáveis é mais importante? Simples: não podemos dizer qual é a mais importante, o modelo pode incluir qualquer uma delas de forma intercambiável.
- **Equações com previsores instáveis**: Descrevi como a multicolinearidade aumenta as variâncias dos coeficientes de regressão, resultando em equações com previsores instáveis. Isso significa que um valor estimado dos coeficientes de regressão (os valores b) será instável de amostra para amostra.

Uma forma de identificar multicolinearidade é examinar a matriz de correlações de todas as variáveis previsoras e ver se alguma se correlaciona de forma bastante alta (por alta queremos dizer acima de 0,80 ou 0,90). Esse é um método prático, mas deixa escapar formas mais sutis de multicolinearidade. Felizmente, o SPSS produz vários diagnósticos de colinearidade, entre eles o FIV (**Fator de Inflação da Variância**). O FIV indica se um previsor tem um relacionamento linear forte com outro(s) previsor(es). Embora não existam regras simples sobre quais valores do FIV devem ser motivo de alerta, Meyers (1990) sugere que 10 é um bom valor a partir do qual podemos nos preocupar. Bowerman e O'Connel (1990) sugerem que se na média o FIV é maior do que 1, a multicolinearidade pode tornar o modelo de regressão tendencioso. Relacionado ao valor FIV existe a estatística **tolerância**, que é o valor inverso de FIV (1/FIV). Assim, valores

Material com direitos autorais

abaixo de 0,10 indicam problemas sérios, embora Menard (1995) sugira que valores abaixo de 0,20 já sejam motivo de preocupação.

Outra medida útil na descoberta de valores de previsores dependentes são os *autovalores da matriz dos produtos cruzados não reduzida e não centrada*, o *índice de condição* e as *proporções da variância*. Essas estatísticas são extremamente complexas e serão abordadas na parte de interpretação das saídas do SPSS (veja a Seção 5.8.5). Se nada disso faz sentido para você, dê uma olhada em Hutcheson e Sofroniou (1999, p. 78-85), que fornece uma explicação bastante clara da multicolinearidade.

5.7 COMO EXECUTAR UMA REGRESSÃO MÚLTIPLA NO SPSS ②

5.7.1 Principais opções ②

Imagine que um executivo de uma gravadora está interessado em estender o modelo da venda de discos para incorporar outras variáveis. Ele decide mensurar duas novas variáveis: (1) o número de vezes que o disco toca na Radio 1 (a maior estação de rádio da Grã-Bretanha) durante a semana anterior ao lançamento (**airplay**) e (2) a atratividade da banda (**attract**). Antes de um disco ser lançado, o executivo anota a quantia gasta em publicidade, o número de vezes que o disco é tocado na rádio na semana anterior ao seu lançamento e a atratividade da banda. Ele faz isso para 200 tipos diferentes de discos (cada um de uma banda diferente). A atratividade é mensurada solicitando a uma amostra aleatória da audiência pretendida que atribua uma nota variando de 0 (insossos) a 10 (maravilhosos) para cada banda. A moda da atratividade atribuída para cada banda foi utilizada na regressão (porque ele estava interessado no que a maioria das pessoas pensava, em vez da opinião média das pessoas).

Esses dados estão no arquivo **Record2.sav** (Discos2.sav) e você deve perceber que cada variável tem sua própria coluna (o mesmo método utilizado para a correlação) e que cada linha representa um disco diferente. Assim, o primeiro disco apresenta £10 269 de gastos em publicidade, vendeu 330 000 copias, foi executado 43 vezes na Radio 1 na semana anterior ao seu lançamento e a maioria das pessoas julgou a banda maravilhosa (Figura 5.10).

O executivo tem pesquisas anteriores indicando que a quantia investida em publicidade é um previsor significativo da venda de discos e, assim, ele deve incluir essa variável primeiro no modelo. Suas novas variáveis (**airplay** e **attract**) devem, dessa forma, entrar no modelo

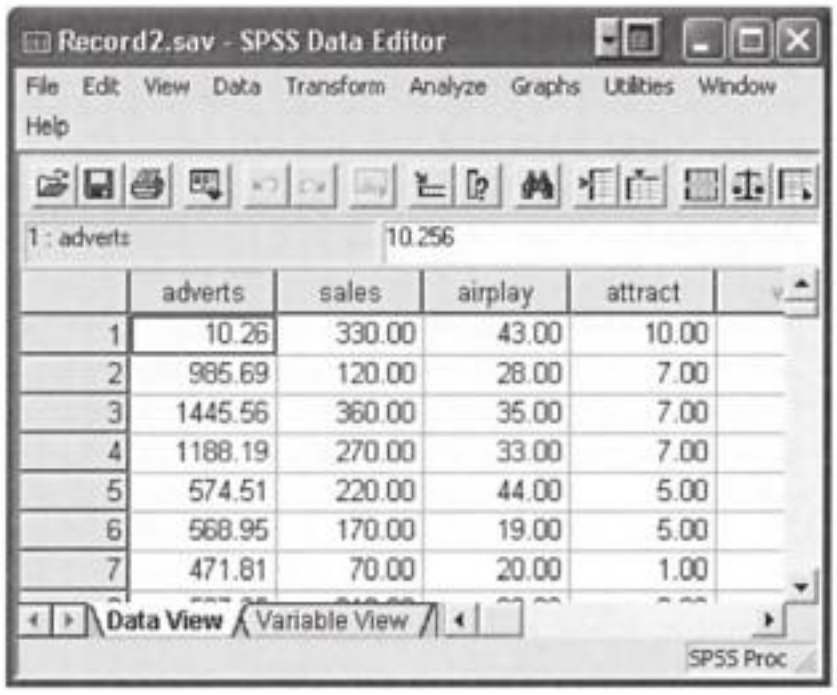
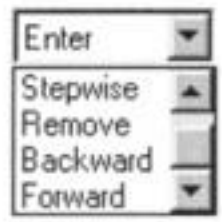


Figura 5.10 Apresentação dos dados para a regressão múltipla.

Material com direitos autorais

após a quantia investida em publicidade. Esse modelo é hierárquico (o pesquisador decide a ordem na qual as variáveis devem ser agregadas ao modelo com base em pesquisas anteriores). Para executar o modelo hierárquico no SPSS, temos que entrar as variáveis em blocos (cada bloco representando um passo na hierarquia). Para chegar à caixa de diálogo principal da regressão (*regression*), você deve clicar no menu *Analyze* (Analisar) e selecionar *Regression* (Regressão) e depois *Linear* (Linear) (**Analyze⇒Regression⇒Linear**). A caixa de diálogo principal é mostrada na Figura 5.11 – ela é a mesma que foi utilizada na regressão simples.

A caixa de diálogo principal é bastante autoexplicativa: existe um espaço para especificar a variável dependente (saída ou resultado) e um espaço para colocar uma ou mais variáveis independentes de entrada (variáveis previsoras). As variáveis no editor de dados são listadas num painel à esquerda. Destaque ou marque a variável *record sales* (vendas de discos) na lista clicando sobre ela e a transfira para a caixa denominada *Dependent* (Dependente) clicando em ▶. Também precisamos especificar a variável previsora para o primeiro bloco. Foi decidido que a quantia aplicada em publicidade deve ser a primeira variável a entrar no modelo (porque pesquisas anteriores indicam que ela é um previsor importante), assim, marque-a na lista e a transfira para o quadro à direita rotulado de *Independent(s)* (Independente(s)) clicando em ▶. Na parte inferior do quadro das variáveis independentes (*Independent(s)*) existe um menu tipo lista suspensa onde é possível especificar o método (*Method*) da regressão (veja a Seção 5.5.3). Você pode selecionar um método diferente de entrada de variáveis para cada bloco clicando em ▼, próximo de onde diz *Method* (Método). A opção por omissão (*default*) é a entrada forçada e essa é a opção que queremos, mas se você estiver executando um trabalho mais exploratório, pode escolher um dos métodos passo a passo (para a frente – *forward*, para trás – *backward*, passo a passo – *stepwise* ou remover – *remove*).



Uma vez especificado o primeiro bloco da hierarquia, precisamos nos mover para o segundo. Para informar ao computador que queremos especificar um novo bloco de va-

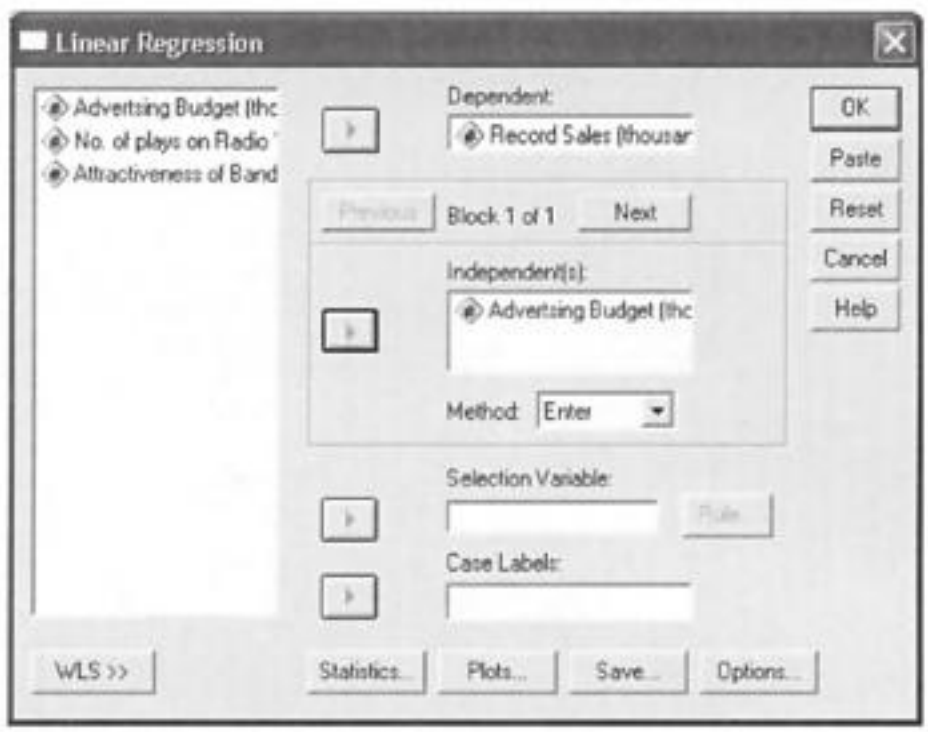


Figura 5.11 Caixa de diálogo principal para o primeiro bloco da regressão múltipla.

Material com direitos autorais

riáveis previsoras você deve clicar em Next (próximo). Esse processo limpa o quadro *Independent(s)* (Independente) de maneira que você pode entrar os novos previsores (você deve notar, também, que acima desse bloco está escrito agora *Block 2 of 2* (Bloco 2 de 2), indicando que você está no segundo bloco dos dois que foram especificados). Decidimos que o segundo bloco deveria conter os dois novos previsores, portanto, você deve clicar em **airplay** e **attract** na lista de variáveis e transferi-las, uma a uma, para o quadro rotulado de *Independent(s)* (Independente(s)) clicando em ▶. A caixa de diálogo fica semelhante a da Figura 5.12. Para se movimentar entre os blocos, utilize os botões Previous (prévio) e Next (próximo) (por exemplo, para voltar para o bloco 1, clique em Previous).

É possível selecionar métodos diferentes de entrada de uma variável para blocos diferentes em uma hierarquia. Assim, embora tivéssemos especificado a entrada forçada para o primeiro bloco, podemos agora especificar um método passo a passo para o segundo. Dado que não temos pesquisas anteriores em relação às variáveis *attractiviness* (atratividade) e *airplay* (número de reproduções na Radio 1), em relação à vendas podemos justificar a escolha pelo método passo a passo para esse bloco. Contudo, em virtude de problemas com o método passo a passo, optarei pelo de entrada forçada para os dois blocos neste exemplo.

5.7.2 Estatísticas ②

Na caixa de diálogo principal da regressão (*regression*), clique em Statistics... (estatísticas) para abrir a caixa de diálogo e selecionar várias opções importantes relacionadas ao modelo (Figura 5.13). Muitas dessas opções estão relacionadas aos parâmetros do modelo; contudo, existem procedimentos disponíveis para verificar a hipótese da não-existência de multicolinearidade (diagnóstico de colinearidade) e independência serial dos erros (Durbin-Watson). Quando você tiver selecionado as estatísticas necessárias (recomendo selecionar todas exceto a matriz de covariâncias como regra geral), clique em Continue (continue) para retornar à caixa de diálogo principal.

- **Estimates** (Estimativas): Essa opção é selecionada por omissão porque ela nos fornece os coeficientes estimados do modelo de regressão (isto é, estimativas dos valores b). A estatística teste e sua significância são fornecidas para cada coeficien-

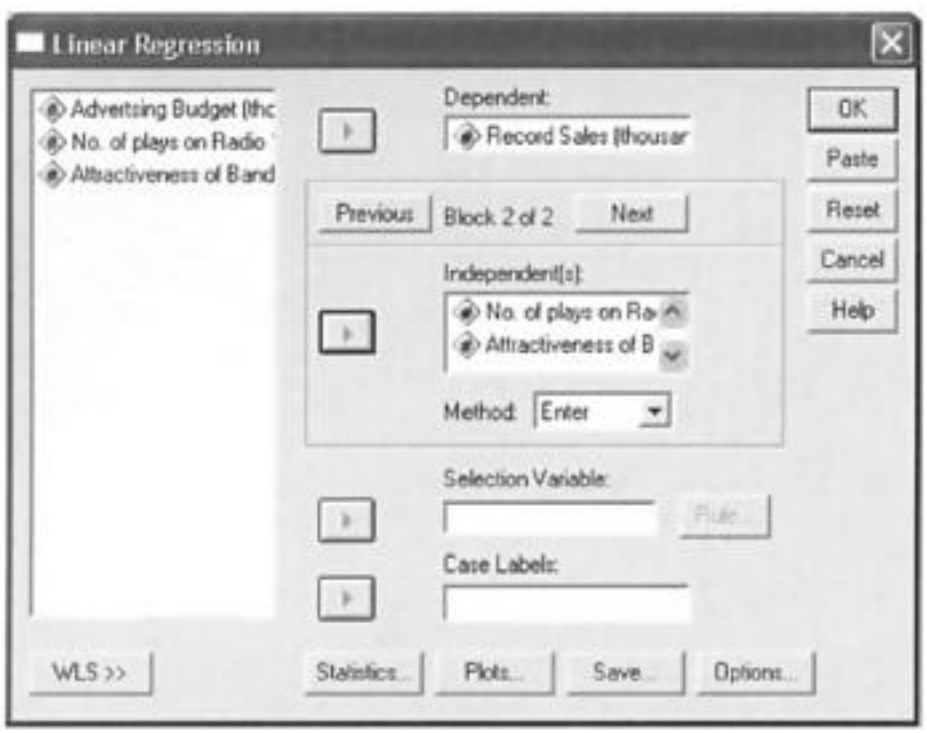


Figura 5.12 Caixa de diálogo principal para o segundo bloco da regressão múltipla.

Material com direitos autorais

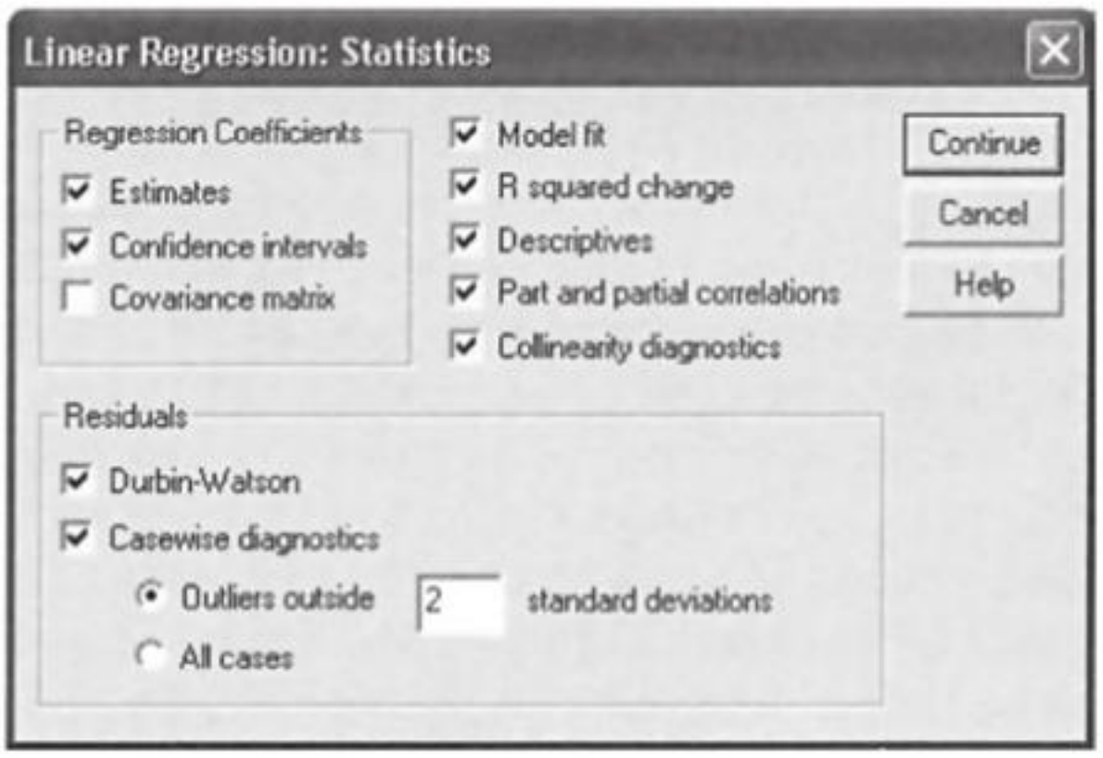


Figura 5.13 Caixa de diálogo *Statistics* (Estatísticas) para a análise de regressão.

te de regressão: um teste *t* é utilizado para verificar se cada b difere significantemente de zero (veja a Seção 5.2.4).

- **Confidence intervals** (Intervalos de confiança): Essa opção, se selecionada, produz intervalos de confiança para um dos coeficientes de regressão não-padronizados. Intervalos de confiança podem ser recursos úteis para avaliar a probabilidade dos coeficientes de regressão na população – irei descrever as interpretações exatas mais tarde.
- **Covariance matrix** (Matriz de covariâncias): Se selecionada, essa opção irá apresentar a matriz de covariâncias, os coeficientes de correlação e as variâncias entre os coeficientes de regressão para cada variável no modelo. Uma matriz de variâncias-covariâncias é produzida com as variâncias apresentadas na diagonal e as covariâncias como elementos fora da diagonal principal. As correlações são apresentadas em uma matriz separada.
- **Model fit** (Aderência do modelo): Essa opção é vital e, portanto, ela é selecionada por omissão. Ela fornece não somente um teste estatístico da habilidade de o modelo prever a variável de saída (o teste F – veja a Seção 5.2.3), mas também o valor de *R* (ou *R* múltiplo), o R^2 correspondente e o R^2 ajustado.
- **R squared change** (Alterações no *R* ao quadrado): Essa opção apresenta as alterações que ocorrem no R^2 resultantes da inclusão de um novo previsor (ou bloco de previsores). Essa medida é uma maneira útil de avaliar a contribuição do novo previsor (ou conjunto de previsores) na explicação da variância das saídas.
- **Descriptives** (Descritivas): Se selecionada, essa opção mostrará uma tabela com a média, o desvio padrão e o número de observações de todas as variáveis incluídas na análise. A matriz de correlações é também apresentada mostrando as correlações entre todas as variáveis e a probabilidade unilateral para cada coeficiente de correlação. Essa opção é extremamente útil porque a matriz de correlação pode ser utilizada para avaliar se os previsores estão inter-relacionados (o que pode ser utilizado para verificar se existe multicolinearidade).
- **Part and partial correlations** (Correlação parcial e por partes): Essa opção produz a correlação de ordem-zero (a correlação de Pearson) entre cada previsor e a saída, controlada para todos os outros previsores do modelo. Ele produz a correlação por partes (ou correlação semiparcial) entre cada previsor e a saída. Essa correlação representa o relacionamento entre cada previsor e a parte da saída que não é explicada pelos outros previsores no modelo. Como tal, ela mede o relacionamento único entre um previsor e a saída (veja a Seção 4.6).

Material com direitos autorais

- **Collinearity diagnostics** (Diagnóstico de colinearidade): Essa opção serve para obter as estatísticas de colinearidade, como VIF, tolerância, autovalores da matriz reduzida e não-centrada dos produtos cruzados e índice de condição e proporção de variâncias (veja a Seção 5.6.2.4).
- **Durbin-Watson:** Essa opção produz a estatística teste de Durbin-Watson, que testa a suposição de independência dos erros. Infelizmente, o SPSS não fornece o valor da significância desse teste, assim, você mesmo deve decidir se o valor é suficientemente diferente de 2 para ser motivo de preocupação (veja a Seção 5.6.2.1).
- **Casewise diagnostics** (Diagnósticos por casos): Essa opção, se selecionada, lista os valores observados da saída, os valores de saída previstos, a diferença entre esses valores (os resíduos) e essa diferença padronizada. Além disso, esses valores podem ser listados para todos os casos ou apenas para casos onde o resíduo padronizado for maior do que 3 (quando o sinal de $\pm$ for ignorado). O critério de valor igual a 3 poderá ser alterado e eu recomendo alterá-lo para 2 por razões que logo ficarão evidentes. Uma tabela resumo de estatísticas residuais indicando o mínimo, o máximo, a média e o desvio padrão tanto dos valores previstos pelo modelo quanto dos resíduos (veja a Seção 5.8.6) também é produzida.

5.7.3 Diagramas da regressão ②

Uma vez que você esteja de volta à caixa de diálogo principal, clique em Plots... (Gráficos) para ativar a caixa de diálogo para os diagramas (*plots*) da regressão, mostrada na Figura 5.14. Essa caixa de diálogo fornece os meios para especificar vários gráficos, que podem auxiliar no estabelecimento da validade de algumas hipóteses da regressão. A maior parte desses gráficos envolvem vários valores residuais (*residual*), que serão descritos com mais detalhes na Seção 5.7.4.

No lado esquerdo da caixa há uma lista de diversas variáveis.

- **DEPENDNT** (a variável de saída).
- ***ZPRED** (*os valores previstos padronizados* da variável dependente com base no modelo). Esses valores são formas padronizadas dos valores previstos pelo modelo.
- ***ZRESID** (*os resíduos padronizados*, ou erros). Esses valores são as diferenças padronizadas entre os dados observados e os valores que o modelo prevê.
- ***DRESID** (*os resíduos excluídos*). Veja a Seção 5.6.1.1 para detalhes.
- ***ADJPRED** (*os valores previstos ajustados*). Veja a Seção 5.6.1.2 para detalhes.
- ***SRESID** (*os resíduos estudentizados*).
- ***SDRESID** (*os resíduos estudentizados excluídos*). Esse valor é o resíduo excluído dividido pelo erro padrão.

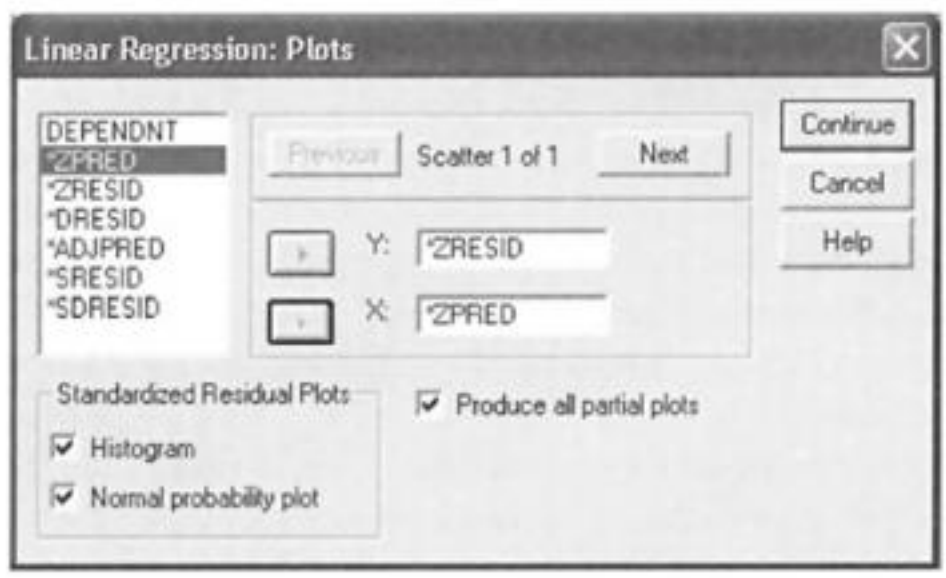


Figura 5.14 Regressão linear: caixa de diálogo gráfica.

As variáveis listadas nessa caixa de diálogo são todas apresentadas sob um título geral de resíduos e foram discutidas em detalhe na Seção 5.6.1.1. Para uma análise básica vale à pena obter os diagramas do *ZRESID (eixo y) contra *ZPRED (eixo x), porque esse gráfico é útil para determinar se as hipóteses de erros aleatórios e homocedasticidade foram satisfeitas. Um diagrama do *SRESID (eixo y) contra *ZPRED (eixo x) mostrará qualquer heterocedasticidade também. Embora muitas vezes esses dois diagramas sejam praticamente idênticos, o segundo é mais sensível em uma situação caso a caso. Para criar esses diagramas, basta selecionar uma variável da lista e transferi-la para o espaço denominado tanto *X* quanto *Y* (referentes aos eixos) clicando em ▶. Quando você tiver selecionado as variáveis para o primeiro diagrama (neste caso, semelhante ao da Figura 5.14) você pode especificar um novo diagrama clicando em Next (próximo). Esse procedimento limpa os espaços em que as variáveis foram especificadas. Se você clicar em Next e quiser retornar ao último diagrama especificado, clique em Previous (anterior). Você pode especificar até nove diagramas.

É possível selecionar a caixa de marcar denominada *Produce all partial plots* (Produza todos os diagramas parciais), que irá determinar diagramas de dispersão dos resíduos da variável de saída e cada um dos previsores quando ambas as variáveis são analisadas separadamente com os previsores restantes. Mesmo que você não tenha entendido a frase anterior, esses diagramas apresentam várias características importantes que faz valer à pena inspecioná-los. Primeiro, o gradiente da linha de regressão entre duas variáveis residuais é equivalente ao coeficiente do previsor na equação de regressão. Assim, qualquer valor atípico óbvio em um diagrama parcial representa um caso que pode ter excessiva influência no coeficiente de regressão de um previsor. Segundo, relacionamentos não-lineares entre um previsor e a variável de saída são mais detectáveis utilizando esses gráficos. Finalmente, eles são formas úteis de detectar colinearidade. Por essas razões, recomendo que eles sejam utilizados.

Existem várias opções para apresentar graficamente os resíduos. Primeiro, você pode selecionar um histograma dos resíduos padronizados (isto é extremamente útil para verificar a hipótese da normalidade dos erros). Segundo, você pode solicitar um diagrama de probabilidade normal, que também informa se os resíduos do modelo são normalmente distribuídos. Quando você tiver selecionado as opções necessárias clique em Continue para voltar para a caixa de diálogo principal da regressão.

5.7.4 Salvando os diagnósticos da regressão ②

Na Seção 5.6, vimos dois tipos de diagnóstico da regressão: aqueles que nos ajudam a avaliar quão bem o modelo se ajusta a nossa amostra e aqueles que nos auxiliam a detectar casos que apresentam uma grande influência no modelo gerado. No SPSS, podemos escolher salvar essas variáveis diagnóstico no editor de dados (assim, o SPSS irá calculá-las e criar novas colunas no editor de dados em que os valores serão apresentados).

Para salvar os diagnósticos da regressão, você precisa clicar em Save... (salvar) na caixa de diálogo principal da regressão. Esse procedimento irá ativar a caixa de diálogo para salvar as variáveis (veja a Figura 5.15). Uma vez que a janela esteja ativa, basta marcar as opções desejadas nos quadrinhos à direita de cada uma. A maioria das opções disponíveis foram explicadas na Seção 5.6, e a Figura 5.15 mostra o que eu considero um conjunto básico de diagnóstico. Versões padronizadas (e estudentizadas) desses diagnósticos são geralmente mais fáceis de interpretar, assim, sugiro selecioná-las em vez das versões não-padronizadas. Uma vez que a regressão tenha sido executada, o SPSS irá criar uma coluna no editor de dados para cada uma das estatísticas solicitadas, e ele tem um conjunto padronizado de nomes dessas variáveis para nomear cada uma. Depois do nome existirá um número que se refere à análise que foi executada. Assim, na primeira vez que a regressão é executada, o conjunto de nomes das variáveis será seguido do valor 1; se você executar uma segunda

Material com direitos autorais

Figura 5.15 Caixa de diálogo para os diagnósticos da regressão.

regressão, ele irá criar um segundo conjunto de variáveis com o nome seguido pelo valor 2, e assim por diante. Os nomes das variáveis no editor de dados são apresentados a seguir. Depois que você selecionar os diagnósticos desejados (clicando nos lugares adequados), clique em Continue para retornar à caixa de diálogo principal da regressão.

- **pre_1:** (valor previsto não-padronizado).
- **zpr_1:** (valor previsto padronizado).
- **adj_1:** (valor previsto ajustado).
- **sep_1:** (erro padrão do valor previsto).
- **res_1:** (resíduo não-padronizado).
- **zre_1:** (resíduo padronizado).
- **sre_1:** (resíduo estudentizado).
- **dre_1:** (resíduo excluído)
- **sdr_1:** (resíduo estudentizado excluído).
- **mah_1:** (distância de Mahalanobis).
- **coo_1:** (distância de Cook).
- **lev_1:** (valor centrado da influência).
- **sdb0_1** (DFBETA padronizado (intercepto)).
- **sdb1_1** (DFBETA padronizado (previsor 1)).
- **sdb2_1:** (DFBETA padronizado (previsor 2)).
- **sdf_1:** DFFIT padronizado.
- **cov_1:** (razão de covariância).

5.7.5 Opções adicionais ②

Como uma etapa final da análise você pode clicar em Opções... para levá-lo à caixa de diálogo Opções (Figura 5.16). O primeiro conjunto de opções permite que você mude os critérios utilizados para entrar as variáveis na regressão passo a passo (*stepwise*). Se você insistir em realizar a regressão passo a passo, provavelmente é melhor manter o critério padrão de uma probabilidade de 0,05. Contudo, você pode tornar esse critério mais restrito (0,01). Existe outra opção para construir um modelo que não inclui a constante (isto é, não apresenta interseção com o eixo Y). Essa opção também deve ser ignorada! Finalmente, você pode selecionar um método para lidar com os pontos de dados omitidos (*missing*). Por padrão, o SPSS exclui casos por linhas (*listwise*), o que significa que se uma pessoa

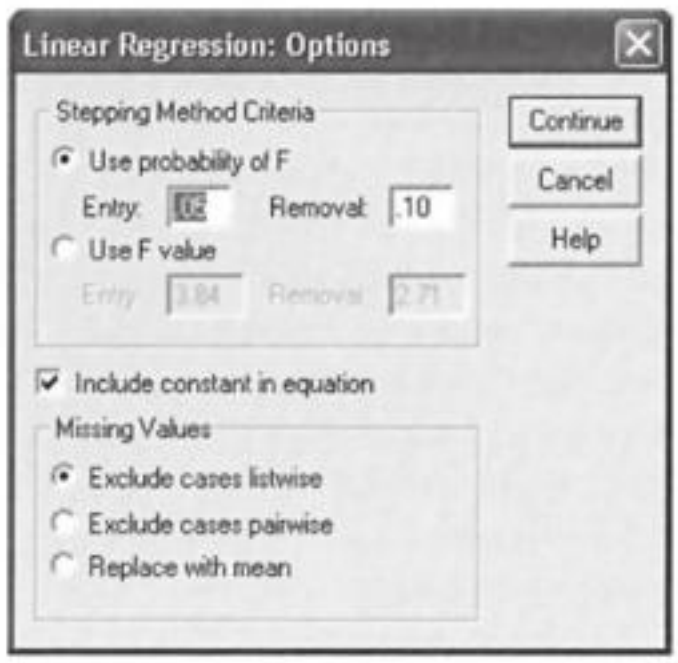


Figura 5.16 Opções para a regressão linear.

tem um valor não conhecido em uma variável, ela será excluída da análise. Assim, por exemplo, se o nosso executivo da gravadora não tem um escore para a atratividade de uma de suas bandas, os dados desse disco não serão utilizados no modelo de regressão. Outra opção é excluir casos em uma base de pares, o que significa que se o participante tem um escore desconhecido para uma variável em particular, seus dados serão excluídos da análise somente em cálculos envolvendo a variável para a qual ele não apresenta valor. Assim, o dado para a banda que não apresenta escore de atratividade seria utilizado para calcular o relacionamento entre valor investido em publicidade, número de vezes que o disco foi executado na Radio 1 e discos vendidos. Contudo, se você fizer isso, muitas das suas variáveis podem não fazer sentido, e talvez você acabe gerando absurdos como um R^2 tanto negativo quanto maior do que 1,0. Portanto, essa não é uma boa opção.

Outra possibilidade é substituir o valor que falta com um valor médio dessa variável e depois incluir o caso na análise (assim, no nosso exemplo a banda teria um valor de atratividade igual à medida da atratividade de todas as bandas). O problema dessa última escolha é que provavelmente o verdadeiro valor do desvio padrão será reduzido (e, ainda mais importante, o do erro padrão). O desvio padrão será diminuído porque para cada caso que for substituído não existirá diferença entre a média e o escore a ele atribuído, enquanto que se o dado tivesse sido coletado para o tal caso ele quase certamente iria apresentar alguma diferença em relação à média. Obviamente, se a amostra é grande e o número de valores desconhecidos é pequeno, esse não é um problema sério. Contudo, se existirem muitos valores desconhecidos, essa escolha é potencialmente perigosa porque erros padrão pequenos terão maior probabilidade de resultar em valores significativos que serão o produto da substituição dos dados e não um efeito verdadeiro. A opção final é utilizar a rotina do SPSS "Análise de valores desconhecidos", destinada para especialistas. Ela faz uso do fato de que se as suas variáveis estão presentes e correlacionadas para muitos casos no arquivo e um valor ocasional estiver faltando, você poderá substituí-lo por estimativas melhores do que a média (Tabachnick e Fidell, 2001, Capítulo 4, descrevem alguns desses procedimentos).

5.8 INTERPRETANDO A REGRESSÃO MÚLTIPLA ②

Uma boa estratégia para adotar com a regressão é medir as variáveis previsoras para as quais existem razões teóricas para esperar que prevejam bem o resultado. Execute a análise de regressão em que todos os previsores são colocados no modelo e examine a saída para ver quais previsores contribuem substancialmente para o modelo prever um resultado. Uma vez que você tenha determinado quais são as variáveis importantes, execute novamente a análise incluindo somente esses previsores e utilize as estimativas dos parâmetros resultantes para definir o modelo de regressão. Se a análise inicial revelar que existem dois ou mais previsores significativos, você pode considerar a execução de uma análise passo a passo para frente (*forward stepwise*) (em vez de uma entrada forçada) a fim de encontrar a contribuição individual de cada previsor.

Gastei um monte de tempo explicando a teoria de base da regressão e de algumas das ferramentas de diagnóstico necessárias para

medir a precisão de um modelo de regressão. É importante lembrar que o SPSS pode parecer bastante esperto, porém, de fato, ele não é. Ele pode, sim, realizar muitos cálculos complexos em questão de segundos, mas ele não consegue controlar a qualidade do modelo gerado – para tanto, é imprescindível um cérebro humano (e de preferência um que tenha sido treinado). O SPSS pode gerar facilmente saídas com base em qualquer tipo de lixo fornecido no editor de dados e ele não julgará julgar os resultados ou indicar se o modelo é generalizável ou mesmo válido. Contudo, o SPSS fornece as estatísticas necessárias para julgar essas coisas e a partir desse ponto nós devemos fazer o trabalho – o que é levemente preocupante (especialmente se o seu cérebro é tão pequeno quanto o meu!).

Tendo selecionado todas as opções relevantes e retornado à caixa de diálogo principal, precisamos clicar em OK para executar a análise. O SPSS exibirá inúmeras quantidades de saídas na janela visualizadora e nós precisamos descobrir como fazer com que essas informações façam sentido.

5.8.1 Descritivas ②

A saída descrita nesta seção é produzida utilizando a opção na caixa de diálogo *statistics* (estatísticas) da regressão linear (veja a Figura 5.13). Para começar, se você selecionou a opção *Descriptives* (Descritivas), o SPSS irá produzir a tabela visualizada na saída do SPSS 5.4. Essa tabela nos fornece a média e o desvio padrão de cada variável do nosso conjunto de dados, assim, sabemos que o número médio de discos vendidos foi de 193.200. Essa tabela não é necessária para interpretar o modelo de regressão, mas é útil como resumo dos dados. Além das estatísticas descritivas, a seleção dessa opção fornece também uma matriz de correlações. A tabela mostra três coisas. Primeiro, é apresentado o valor do coeficiente de correlação de Pearson entre cada par de variáveis (por exemplo, podemos ver que o gasto em publicidade tem uma grande correlação positiva com a venda de discos, $R = 0{,}578$). Segundo, é mostrada a significância unilateral de cada correlação (por exemplo, a correlação acima é significativa, $p < 0{,}001$). Finalmente, é mostrado o número de casos que contribui com cada correlação ($N = 200$).

Você deve ter notado que ao longo da diagonal da matriz os valores para os coeficientes de correlação são todos iguais a 1,00 (isto é, uma correlação positiva perfeita). O motivo disso é que esses valores representam a correlação de cada uma das variáveis com ela mesma, assim, obviamente os valores resultantes são iguais a 1. A matriz de correlações é extremamente útil para fornecer uma ideia aproximada do relacionamento entre os previsores e a variável de saída e para um primeiro exame da multicolinearidade. Se não existir multicolinearidade nos dados, não deve existir valores de correlação substanciais ($R > 0{,}90$) entre os previsores.

Se olharmos apenas para os previsores (ignorando as vendas de discos), a correlação mais alta é entre a atratividade da banda e a quantidade de execução significativa ao nível de 0,01 ($R = 0{,}182$ e $p = 0{,}005$). A despeito da significância dessa correlação, o coeficiente é pequeno e, assim, parece que os previsores estão medindo coisas diferentes (não existe colinearidade). Podemos ver, também, que de todos os previsores, o número de execuções na Radio 1 é o que melhor se correlaciona com a saída ($R = 0{,}599$, $p < 0{,}001$); portanto, é provável que essa variável seja a melhor para prever a venda de discos.

Dica da Samanta Ferrinho

Use as estatísticas descritivas para verificar na matriz de correlações a multicolinearidade; isto é, os previsores altamente correlacionados entre si, $R > 0{,}90$.

Material com direitos autorais

Saída 5.4 do SPSS Estatísticas descritivas para a análise de regressão

Descriptive Statistics **(Estatística Descritiva)**

	Mean (Média)	*Std. Deviation* (Desvio padrão)	*N*
Record Sales (thousands) (Vendas de discos) (milhares)	193.2000	80.6990	200
Advertising Budget (Thousands of pounds) Verba de publicidade (Milhares de libras)	614.4123	485.6552	200
No. Of plays on Radio 1 per week (Número de execuções na Radio 1 por semana)	27.5000	12.2696	200
Attractiveness of Band (Atratividade da Banda)	6.7700	1.3953	200

Correlations **(Correlações)**

		Record Sales (thousands) (Vendas de discos) (milhares)	*Advertising Budger* (Verba de publicidade)	*No. Od plays on Radio 1 per week* (Número de execuções na Radio 1 por semana)	*Attractiveness of Band* (Atratividade da Banda)
Pearson Correlation (Correlação de Pearson)	*Record Sales (thousands)* (Vendas de discos) (milhares)	1.000	0.578	0.599	0.326
	Advertising Budget (Thousands of pounds) Verba de publicidade (Milhares de libras)	0.578	1.000	0.102	0.081
	No. of plays on Radio 1 per week (Número de execuções na Radio 1 por semana)	0.599	0.102	1.000	0.182
	Attractiveness of Band (Atratividade da Banda)	0.326	0.081	0.182	1.000
Sig. (1-tailed) (Sig.) (Unilateral)	*Record Sales (thousands)* (Vendas de discos) (milhares)	.	0.000	0.000	0.000
	Advertising Budget (Thousands of pounds) Verba de publicidade (Milhares de libras)	0.000	.	0.076	0.128
	No. of plays on Radio 1 per week (Número de execuções na Radio 1 por semana)	0.000	0.076	.	0.005
	Attractiveness of Band (Atratividade da Banda)	0.000	0.128	0.005	.
N	*Record Sales (thousands)* (Vendas de discos) (milhares)	200	200	200	200
	Advertising Budget (Thousands of pounds) Verba de publicidade (Milhares de libras)	200	200	200	200
	No. of plays on Radio 1 per week (Número de execuções na Radio 1 por semana)	200	200	200	200
	Attractiveness of Band (Atratividade da Banda)	200	200	200	200

5.8.2 Resumo do modelo ②

A próxima seção da saída descreve o modelo globalmente (ela informa se o modelo é eficaz em prever a venda de discos). Lembre que escolhemos o método hierárquico e, portanto, cada conjunto de estatísticas-resumo é repetido para cada estágio na

Material com direitos autorais

Saída 5.5 do SPSS Resumo do modelo de regressão

***Model Summary*[c] (Resumo do modelo)**

Model (Modelo)	*R*	*R Square* (R Quadrado)	*Adjusted R Square* (R Quadrado Ajustado)	*Std. Error of the Estimate* (Erro Padrão da Estimativa)	*Change Statistics* (Estatísticas de Mudanças)					*Durbin-Watson*
					R Square Change (Mudança no R Quadrado)	*F Change* (Mudança no F)	*df1* (gl1)	*df2* (gl2)	*Sig. F. Change* (Mudança na Sig. do F)	
1	0.578[a]	0.335	0.331	65.9914	0.335	99.587	1	198	0.000	
2	0.815[b]	0.665	0.660	47.0873	0.330	96.447	2	196	0.000	1.950

a. Previsores: (constante), Verba publicitária (em milhares de libras).
b. Previsores: (constante), Verba publicitária (em milhares de libras), Atratividade da Banda, Número de execuções por semana na Radio 1.
c. Variável dependente: Vendas de discos (em milhares).

hierarquia. Na Saída do SPSS 5.5, você deve notar que existem dois modelos. O Modelo 1 se refere ao primeiro estágio da hierarquia quando apenas a verba de publicidade é utilizada como um previsor. O Modelo 2 se refere a quando todos os três previsores são utilizados. A Saída do SPSS 5.5 é o *resumo do modelo* (model summary) e essa tabela foi gerada marcando-se a opção *Model fit* (Ajustar modelo). Essa opção é selecionada por omissão no SPSS porque ela nos fornece algumas informações importantes sobre o modelo: os valores de R, R^2 e o R^2 ajustado. Se as opções *R squared change* (Mudanças no R quadrado) e *Durbin Watson* forem selecionadas, esses valores também serão incluídos (se eles não forem selecionados, você terá uma tabela menor).

A tabela resumo é mostrada na Saída do SPSS 5.5, e você notará que ela informa qual era a variável dependente (outcome) e quais eram os previsores em cada um dos dois modelos. Na coluna denominada R estão os valores dos coeficientes de correlação múltipla entre os previsores e a saída. Quando somente a verba de publicidade é utilizada como um previsor, existe uma correlação simples entre os valores investidos em publicidade e a venda de discos (0,578). Na verdade todas as estatísticas do modelo 1 são as mesmas do modelo de regressão simples visto anteriormente (Seção 5.4). A próxima coluna fornece o valor de R^2, que nós já conhecemos como uma medida de quanta variabilidade da saída pode ser debitada aos previsores. Para o primeiro modelo esse valor é 0,335, o que significa que a verba de publicidade é responsável por 33,5% da variação nas vendas de discos. Contudo, quando os outros dois previsores são também incluídos (modelo 2), esse valor aumenta para 0,665 ou 66,5% da variância na venda de discos. Dessa forma, se a publicidade é responsável por 33,5%, podemos verificar que a atratividade e o número de execuções são responsáveis por mais 33%.[9] Assim, a inclusão dos dois novos previsores explicou boa parte da variação das vendas de discos.

O R^2 ajustado fornece uma noção de quão bem nosso modelo generaliza, e idealmente nós gostaríamos que esse valor fosse igual ao, ou muito próximo do, valor do R^2. Nesse exemplo, a diferença para o modelo final é pequena (de fato, a diferença entre os valores é $0{,}665 - 0{,}660 = 0{,}005$ ou aproximadamente 0,5%). Isso significa que se o nosso modelo fosse derivado da população em vez de uma amostra, ele explicaria aproximadamente 0,5% menos da variância da saída. Estudantes avançados podem querer aplicar a fórmula de Stein do R^2 para ter uma ideia de quão provável seria esse valor entre diferentes amostras. A fórmula de Stein foi apresentada na equação

[9] Isto é, 33% = 66,5% − 33,5% (esse valor é o *R Square Change* (alteração no R quadrado) apresentado na tabela).

Material com direitos autorais

(5.11) e pode ser aplicada pela substituição de **n** pelo tamanho da amostra (200) e k com o número de previsores (3):

$$R^2 \text{ ajustado} = 1 - \left[\left(\frac{200-1}{200-3-1}\right)\left(\frac{200-2}{200-3-2}\right)\left(\frac{200+1}{200}\right)\right](1-0{,}665)$$
$$= 1 - [(1{,}015)(1{,}015)(1{,}005)](0{,}335)$$
$$= 1 - 0{,}347$$
$$= 0{,}653$$

Esse valor é muito semelhante ao valor do R^2 observado, de 0,665, indicando que a validação cruzada do modelo é muito boa.

As estatísticas de mudança são fornecidas somente se solicitadas e elas nos informam se a mudança no R^2 é significativa. A significância do R^2 pode ser testada utilizando a razão F, e esse F é calculado a partir da seguinte equação (na qual N é o número de casos ou participantes e k é o número de previsores no modelo):

$$F = \frac{(N-k-1)R^2}{k(1-R^2)}$$

Na saída do SPSS 5.5, a mudança nesse F é relatada para cada um dos blocos da hierarquia. Assim, o modelo 1 causa uma mudança no R^2 de zero para 0,335, e essa mudança na quantidade de variância explicada dá origem a uma razão F de 99,587, que é significativa com uma probabilidade menor do que 0,001. Tenha em mente que para esse primeiro modelo temos apenas um previsor (assim, $k = 1$) e 200 casos (N = 200). Esse F é obtido da equação anterior:[10]

$$F_{\text{Model 1}} = \frac{(200-1-1)0{,}334648}{1(1-0{,}334648)} = 99{,}587$$

A adição de novos previsores (modelo 2) causa um aumento do R^2 por 0,330 (veja acima). Podemos calcular a razão F para essa mudança utilizando a mesma equação, mas porque estamos avaliando as mudanças no modelo, utilizaremos essas mudanças, R^2_{Change}, e o R^2 para o novo modelo (modelo 2; nesse caso, vamos chamá-lo de R_2^2). Também usamos a mudança no número de previsores, k_{Change} (o modelo 1 tem um previsor e o modelo 2 tem três previsores, assim, a mudança no número de previsores é 3 – 1 = 2), e o número de previsores no novo modelo, k_2 (porque estamos interessados no modelo 2). Novamente, se utilizarmos algumas casas decimais a mais do que as que a Tabela do SPSS apresenta, iremos obter aproximadamente a mesma resposta obtida pelo SPSS:

$$F_{\text{Change}} = \frac{(N-k_2-1)R^2_{\text{Change}}}{k_{\text{Change}}(1-R_2^2)}$$
$$= \frac{(200-3-1)\times 0{,}330}{2(1-0{,}664668)}$$
$$= 96{,}44$$

Desse modo, a mudança na variância que pode ser explicada fornece uma razão F de 96,44, que é novamente significativo ($p < 0{,}001$). As estatísticas de mudança, portanto, nos informam sobre as diferenças que ocorrem quando adicionamos novos previsores ao modelo.

Finalmente, se você solicitou a estatística de *Durbin-Watson*, ela será encontrada na última coluna da tabela da saída 5.5 do SPSS. Essa estatística nos informa se a hipótese de independência dos erros é satisfeita (veja a Seção 5.6.2.1). Como uma regra conservadora, sugiro que valores menores do que 1 ou maiores do que 3 devem, definitivamente, ser motivos de preocupação (embora eu o aconselhe a determinar valores precisos para a situação de interesse). Quanto mais próximo de 2 o valor estiver, melhor; para esses dados o valor

[10] Para obter o mesmo valor que o SPSS, precisamos utilizar o valor exato do R^2, que é, 0,3346480676231 (se você não acredita dê um clique duplo na tabela de saída do SPSS que apresenta esse valor, depois dê um clique duplo na célula da tabela contendo o valor do R^2 e você verá que 0,335 torna-se o valor que foi digitado acima!).

Material com direitos autorais

é 1,950, bem perto de 2, ou seja, a hipótese quase com certeza está satisfeita.

A saída 5.6 do SPSS mostra a próxima parte da saída, que contém uma análise de variância (ANOVA) que testa se o modelo é significativamente melhor para prever a saída do que utilizar a média como um "bom palpite". Especificamente, F representa a razão de melhoria na previsão que resulta do ajuste do modelo em comparação com a imprecisão que ainda existe no modelo. Somos informados do valor da soma dos quadrados do modelo (esse valor é o SS_M da Seção 5.2.3 e representa a melhoria na previsão resultante do ajuste de uma linha aos dados em vez de utilizar a média como uma estimativa das saídas). Recebemos ainda o valor da soma dos quadrados dos resíduos (esse valor é o SS_R da Seção 5.2.3 e representa a diferença total entre o modelo e os dados observados). Somos apresentados ainda aos graus de liberdade (gl) para cada termo. No caso de melhoria ocasionada pelo modelo, esse valor é igual ao número de previsores (um para o primeiro modelo e três para o segundo), e para o SS_R ele é o número de observações (200) menos o número de coeficientes no modelo de regressão. O primeiro modelo tem dois coeficientes (um para a variável previsora e outro para a constante), enquanto que o segundo tem quatro (um para cada variável

Saída 5.6 do SPSS

***ANOVA*[c] (Análise de Variância)**

	Model (Modelo)	*Sum of squares* (Soma dos quadrados)	*df* (gl)	*Mean Square* (Média dos quadrados)	*F*	*Sig.*
1	*Regression* (Regressão)	433687.833	1	433687.833	99.587	0.00[a]
	Residual (Resíduo)	862264.167	198	4354.870		
	Total (total)	1295952.0	199			
2	*Regression* (Regressão)	861377.418	3	287125.806	129.498	0.00[b]
	Residual (Resíduo)	434574.582	196	2217.217		
	Total (total)	1295952.0	199			

a. Predictors: (Constant), Advertising Budget (thousands of pounds) (Previsores: (constante), Verba publicitária (em milhares de libras))

b. Predictors: (Constant), Advertising Budget (thousands of pounds), Attractiveness of Band, No. of plays on Radio 1 per week (Previsores: (constante), Verba publicitária (em milhares de libras)), Atratividade da Banda, Número de execuções por semana na Radio 1.

c. Dependent variable: Record Sales (thousands) (Variável dependente: Vendas de discos (em milhares))

Dica da Samanta Ferrinho

O ajuste do modelo de regressão pode ser avaliado utilizando o **resumo do modelo (Model Summary)** e as tabelas da ANOVA do SPSS. Um exame do R^2 informará a proporção da variância que é explicada pelo modelo. Se você tiver feito uma regressão hierárquica, poderá avaliar a melhoria do modelo em cada estágio da análise analisando as mudanças no R^2 e se essa mudança é significativa (procure por valores menores do que 0,05 na coluna denominada *Sig. F Change* (Mudança na significância do F)). A ANOVA nos informa também se o modelo adere de forma significativa aos dados (procure por valores menores do que 0,05 na coluna denominada *Sig.*). Finalmente, existe uma suposição de que os erros na regressão são independentes; essa suposição é provavelmente satisfeita se a estatística de Durbin-Watson estiver próxima de 2 (e entre 1 e 3).

Material com direitos autorais

previsora e um para a constante). Portanto, o modelo 1 tem 198 graus de liberdade, ao passo que o modelo 2 tem 196. A média da soma dos quadrados (MS) é então calculada para cada termo pela divisão do SS pelo gl. A razão *F* é calculada dividindo-se a melhoria no aumento médio de previsão do modelo (MS_M) pela diferença média entre o modelo e os dados observados (MS_R). Se a melhoria devido ao ajuste do modelo de regressão é muito maior do que a variação no interior do modelo, então o valor de *F* será maior do que 1 e o SPSS calcula a probabilidade exata de obter o valor de F por acaso. Para o modelo inicial, a razão *F* é 99,587, o que é absolutamente improvável de ter acontecido por acaso ($p < 0{,}001$). Para o segundo modelo, o valor de *F* é ainda maior (129,498), o que é também muito significativo ($p < 0{,}001$). Esses resultados significam que o modelo inicial melhorou de forma significativa nossa habilidade de prever um valor da variável de saída, mas o novo modelo (com os previsores adicionais) foi ainda melhor (porque o valor de *F* é ainda mais significativo).

5.8.3 Parâmetros do modelo ②

Até agora verificamos várias estatísticas resumo que nos informam se o modelo melhorou ou não a nossa habilidade de prever valores da variável de saída. A próxima parte da saída está relacionada com os parâmetros do modelo. A saída 5.7 do SPSS mostra os parâmetros do modelo para os dois passos da hierarquia. O primeiro passo da hierarquia foi incluir a verba publicitária (como foi feito na regressão simples anteriormente neste capítulo) e, assim, os parâmetros do modelo são idênticos aos parâmetros obtidos na Saída 5.3 do SPSS. Portanto, estamos interessados somente nos parâmetros do modelo final (no qual todos os previsores foram incluídos). O formato das tabelas dos coeficientes dependerá das opções selecionadas. Os intervalos de confiança para os valores *b*, o diagnóstico de colinearidade e as correlações por parte e parciais serão apresentadas somente se forem selecionas na caixa de diálogo conforme apresentado na Figura 5.3.

Lembre que de regressão múltipla o modelo toma a forma da equação (5.9) e nessa equação existem várias quantidades desconhecidas (os valores *b*). A primeira parte da tabela apresenta estimativas dos valores b e esses valores indicam a contribuição individual de cada previsor para o modelo. Se substituirmos os valores b na equação (5.9) veremos que é possível definir o modelo como na equação (5.12):

$$\begin{aligned}\text{Vendas}_i &= b_0 + b_1\text{publicidade}_i + b_2\text{execuções}_i \\ &\quad + b_3\text{atratividade}_i \\ &= -26{,}61 + (0{,}08\text{publicidade}_i) \\ &\quad + (3{,}7\text{execuções}_i) \\ &\quad + (11{,}09\text{atratividade}_i) \qquad (5.12)\end{aligned}$$

Os valores b nos informam sobre o relacionamento entre a venda de discos e cada previsor. Se o valor é positivo, podemos dizer que existe um relacionamento positivo entre o previsor e a saída enquanto que um coeficiente negativo representa um relacionamento negativo. Para esses dados, todos os três previsores apresentam valores b positivos, indicando relacionamentos positivos. Assim, à medida que a verba publicitária aumenta, a venda de discos aumenta; se o número de execuções no rádio aumenta, a venda aumenta, e finalmente se a atratividade da banda aumenta, mais discos serão vendidos. Os valores b nos informam mais do que isso; eles nos informam em que grau cada previsor afeta a saída *se todos os demais previsores forem mantidos constantes.*

- **Verba publicitária (Advertising budget)** ($b = 0{,}085$): Esse valor indica que se a verba publicitária aumentar em uma unidade, a venda de discos irá aumentar 0,085 unidades. Ambas as variáveis foram mensuradas em milhares, portanto, para cada £1000,00 a mais gastas em publicidade, 0,085 milhares de discos (85 discos) serão vendidos. Essa interpretação é verdadeira somente se os efeitos da atratividade da banda e o número de execuções forem mantidos constantes.
- **Número de execuções (airplay)** (b = 3,367): Esse valor indica que se o número de vezes que o disco toca na radio na

Material com direitos autorais

Saída 5.7 do SPSS Coeficientes do modelo de regressão[11]

***Coeficients*[a] (Coeficientes)**

Model (Modelo)		*Unstandardized Coefficients* (Coeficientes não-padronizados)		*Standardized Coefficients* (Coeficientes padronizados)	*t*	*Sig.*	*95% Confidence Interval for B* (Intervalo de Confiança de 95% para B)	
		B	*Std. Error* (Erro Padrão)	*Beta*			*Lower Bond* (Limite Inferior)	*Upper bond* (Limite Superior)
1	*(Constant)* (Constante)	134.140	7.537		17.799	0.000	119.278	149.002
	Advertising Budget (thousands of pounds) (Investimento em Publicidade) (em milhares de libras)	9.612E–02	0.10	0.578	9.979	0.000	0.077	0.115
2	*(Constant)* (Constante)	–26.613	17.350		–1.534	0.127	–60.830	7.604
	Advertising Budget (thousands of pounds) (Investimento em Publicidade) (em milhares de libras)	8.488E–02	0.007	0.511	12.261	0.000	0.071	0.099
	No. Of plays on Radio 1 per week (Número de execuções na Radio 1 por semana)	3.367	0.278	0.512	12.123	0.000	2.820	3.915
	Attractiveness of Band (Atratividade da Banda)	11.086	2.438	0.192	4.548	0.000	6.279	15.894

a. Variável dependente: vendas de discos (em milhares).

***Coeficients*[a] (Coeficientes)**

Model (Modelo)		*Correlations* (Correlações)			*Collinearity Statistics* (Estatísticas de Colinearidade)	
		Zero-order (De ordem zero)	*Partial* (Parcial)	*Part* (Parte)	*Tolerance* (Tolerância)	*VIF*
1	*Advertising Budget (thousands of pounds)* (Investimento em Publicidade) (em milhares de libras)	0.578	0.578	0.578	1.000	1.000
2	*Advertising Budget (thousands of pounds)* (Investimento em Publicidade) (em milhares de libras)	0.578	0.659	0.507	0.986	1.015
	No. Of plays on Radio 1 per week (Número de execuções na Radio 1 por semana)	0.599	0.655	0.501	0.959	1.043
	Attractiveness of Band (Atratividade da Banda)	0.326	0.309	0.188	0.963	1.038

a. Variável dependente: vendas de discos (em milhares).

[11] Para poupar sua visão, dividi esta parte da saída em duas tabelas; contudo, ela deverá aparecer como uma única e longa tabela no visualizador de saídas do SPSS.

Material com direitos autorais

venda de discos irá aumentar 0,511 desvios padrão. O desvio padrão da venda de discos é 80699 e, assim, teremos uma mudança de 41240 vendas (0,511.80699). Portanto, para cada £485655 a mais gasto em publicidade, uma venda extra de discos de 41240 unidades será obtida. Essa interpretação é verdadeira somente se os efeitos da atratividade da banda e o número de execuções no rádio permanecerem constantes.

- **Número de execuções (airplay)** (β *padronizado* = 0,512): Esse valor indica que se o número de vezes que o disco toca na radio na semana anterior ao lançamento aumenta em um desvio padrão (12,27), as vendas aumentarão em 0,512 desvios padrão. O desvio padrão da venda de discos é 80699 e, desse modo, uma mudança de 41320 nas vendas poderá ocorrer (0,512.80699). Portanto, se a Radio 1 toca a música um número extra de 12,27 vezes na semana anterior ao lançamento, uma venda extra de 41320 discos é esperada. Essa interpretação é verdadeira somente se a verba publicitária e a atratividade da banda permanecerem constantes.
- **Atratividade (Atractiveness)** ((β *padronizado* = 0,192): Esse valor indica que se a atratividade da banda aumenta um desvio padrão (1,40 unidades), pode-se esperar 0,192 desvios padrão de unidades adicionais vendidas. Isso representa uma mudança de 15490 discos (0,192.80699). Portanto, uma banda com uma atratividade de 1,40 superior a outra banda pode esperar uma venda adicional de 15490 discos. Essa interpretação somente será verdadeira se a verba publicitária e o número de execuções permanecerem constantes.

Imagine que coletamos 100 amostras de dados e mensuramos as mesmas variáveis do nosso modelo atual. Para cada amostra, podemos criar um modelo de regressão para representar os dados. Se o modelo é fidedigno, esperamos encontrar parâmetros bastante semelhantes em ambos. Portanto, cada amostra deve produzir aproximadamente os mesmos valores b. O intervalo de confiança para os valores β não-padronizados são limites construídos tais que em 95% dessas amostras esses limites irão conter os verdadeiros valores de *b* (veja a Seção 1.6.2). Portanto, se tivermos coletado 100 amostras e calculado os intervalos de confiança para b, diremos que temos uma confiança de 95% de que esses intervalos conterão os verdadeiros valores dos coeficientes *b*. Assim, podemos garantir que o intervalo de confiança que construímos para essas amostras conterá o verdadeiro valor de b populacional. Isso acontecendo, um bom modelo apresentará intervalos de confiança pequenos, indicando que os valores de b nessa amostra estão próximos do verdadeiro valor de b na população. O sinal (positivo ou negativo) dos valores b nos informa sobre a direção do relacionamento entre o previsor e o resultado. Desse modo, podemos esperar que um modelo muito ruim apresente intervalos contendo o zero, indicando que em algumas amostras o previsor tem um relacionamento positivo com o resultado, ao passo que em outras ele tem um relacionamento positivo. Nesse modelo, os dois melhores previsores (verba publicitária e número de execuções) apresentam intervalos bastante estreitos, indicando que as estimativas para o modelo atual são provavelmente representativas dos valores populacionais. O intervalo para a atratividade é mais largo (mas ainda não contém o zero), indicando que o parâmetro para essa variável é menos representativo, mas ainda assim significativo.

Se você solicitou as correlações por parte (ou parciais), elas irão aparecer na saída em colunas separadas da tabela. As correlações de ordem zero são os coeficientes de correlação simples de Pearson (e correspondem aos valores na Saída 5.4 do SPSS). As correlações parciais representam o relacionamento entre cada previsor e a variável de saída, controlados os efeitos dos demais previsores. A correlação por partes representa o relacionamento entre cada previsor e a saída, controlado o efeito que as outras duas variáveis apresentam na variável de saída. De fato, essas correlações por partes representam o relacionamento

Material com direitos autorais

Dica da Samanta Ferrinho

A contribuição individual de cada variável para o modelo de regressão pode ser encontrada na tabela dos **coeficientes** do SPSS. Se você fez uma regressão hierárquica, olhe para os valores do modelo final. Você pode ver se cada variável previsora fez uma contribuição significativa para prever o resultado olhando para a coluna rotulada como *Sig*. (valores menores do que 0,05 são significativos). Você deve também observar os valores β padronizados porque eles informam sobre a importância de cada previsor (quanto maior o valor absoluto, melhor). Os valores da tolerância e do FIV também serão úteis mais tarde, portanto, tome nota deles!

único que cada previsor tem com a variável de saída. Se você optar por executar uma regressão passo a passo (*stepwise*), verá que a variável da entrada é baseada inicialmente na variável com a maior correlação de ordem zero e depois com a correlação por parte com as variáveis restantes. Portanto, o número de execuções será adicionado primeiro (em virtude da maior correlação de ordem zero), depois a verba publicitária (porque sua correlação por partes é maior do que a atratividade) e finalmente a atratividade. Tente executar uma regressão passo a passo para frente (*forward stepwise*) nesses dados para ver se eu estou certo! Finalmente, iremos fornecer detalhes sobre as estatísticas de colinearidade, mas elas serão discutidas na Seção 5.8.5.

5.8.4 Variáveis excluídas ②

A cada estágio da análise de regressão, o SPSS fornece um resumo de qualquer variável que ainda não foi adicionada ao modelo. No modelo hierárquico, esse resumo apresenta detalhes das variáveis que foram especificadas para entrar no modelo em passos subsequentes, e na regressão passo a passo essa tabela (Saída 5.8 do SPSS) contém resumos das variáveis que o SPSS determinou que devem entrar no modelo. Para esse exemplo, existe um resumo das variáveis excluídas no primeiro estágio da hierarquia (não existe resumo para o segundo estágio porque todos os previsores estão no modelo). O resumo fornece uma estimativa de cada valor beta do previsor se ele entrar na equação nesse ponto e fornece um

Saída 5.8 do SPSS

***Excluded Variables*[b] (Variáveis Excluídas)**

Model (Modelo)	*Beta In*	*t*	*Sig.*	*Partial Correlations* (Correlações)	*Collinearity Statistics* (Estatísticas de Colinearidade)		
					Tolerance (Tolerância)	*FIV*	*Minimum Tolerance* (Tolerância Mínima)
1 *No. Of plays on Radio 1 per week* (Número de execuções na Radio 1 por semana)	0.546[a]	12.513	0.000	0.665	0.990	1.010	0.990
Attractiveness of Band (Atratividade da Banda)	0.281[a]	5.136	0.000	0.344	0.993	1.007	0.993

a. Previsores no modelo: (constante), Verba publicitária (em milhares de libras).
b. Variável dependente: Venda de discos (em milhares).

Material com direitos autorais

teste *t* para esse valor. Na regressão passo a passo (*stepwise*), o SPSS entra o previsor com a estatística *t* mais alta e continua entrando previsores até que não reste nenhum com estatística *t* que tenha valor significativo inferior a 0,05. A correlação parcial também fornece algumas indicações sobre qual contribuição (se existir alguma) um previsor excluído faria se ele tivesse entrado no modelo.

5.8.5 Avaliando as hipóteses de não-multicolinearidade ②

A Saída 5.7 do SPSS fornece algumas medidas para verificar a colinearidade nos dados. Especificamente, ela fornece as estatísticas FIV e tolerância (com a tolerância sendo 1 dividido pela FIV). Existem algumas observações da Seção 5.6.2.3 que podem ser aplicadas aqui:

- Se a maior FIV for maior do que 10, então existe motivo para preocupação (Myers, 1990; Bowerman e O'Connell, 1990).
- Se a FIV média é substancialmente maior do que 1, então a regressão pode ser tendenciosa (Bowerman e O'Connell, 1990).
- Tolerância abaixo de 0,10 indica problemas sérios.
- Tolerância acima de 0,20 indica um problema em potencial (Menard, 1995).

Para o modelo atual, os valores FIV estão todos abaixo de 10 e as tolerâncias todas acima de 0,20; portanto, podemos concluir com segurança que não existe colinearidade dentro dos nossos dados. Para calcular a FIV média, simplesmente adicionamos os valores para cada previsor e dividimos pelo número de previsores (k):

$$\overline{\text{FIV}} = \frac{\sum_{i=1}^{k} \text{FIV}_i}{k} = \frac{1{,}015 + 1{,}043 + 1{,}038}{3} = 1{,}032$$

A FIV média está muito próxima de 1 e confirma que a colinearidade não é um problema para esse modelo. O SPSS também produz uma tabela de autovalores da matriz dos produtos cruzados, índices de condição e proporção de variâncias. Existe uma longa discussão, e um exemplo, sobre colinearidade na Seção 6.8 e é apresentada a forma como detectá-la utilizando proporções de variâncias. Portanto, vou me limitar agora em dizer que procuraremos por grandes proporções de variâncias em um mesmo *pequeno* autovalor. Assim, na saída 5.9 do SPSS olhamos para as últimas linhas inferiores da tabela (esses são os pequenos autovalores) e procuramos por variáveis que tenham proporções de alta variância para esse autovalor. As

Saída 5.9 do SPSS

***Collinearity Diagnostics*[a] (Diagnósticos de Colinearidade)**

Model (Modelo)	*Dimension* (Dimensão)	*Eingenvalue* (Autovalor)	*Condition Index* (Índice de Condição)	*Variance Proportions* (Proporções da Variância)			
				(Constant) (Constante)	*Advertising Budget (thousands of pounds)* (Verba publicitária (em milhares de libras))	*No. Of plays on Radio 1 per week* (Execuções na Radio 1 por semana)	*Attractiveness of Band* (Atratividade da Banda)
1	1	1.785	1.000	0.11	0.11		
	2	0.215	2.883	0.89	0.89		
2	1	3.562	1.000	0.00	0.02	0.01	0.00
	2	0.308	3.401	0.01	0.96	0.05	0.01
	3	0.109	5.704	0.05	0.02	0.93	0.07
	4	2.039E–02	13.219	0.94	0.00	0.00	0.92

a. Dependent Variable: Record Sales (thousands) (Variável dependente: Venda de discos (em milhares))

Material com direitos autorais

proporções de variâncias variam entre 0 e 1 e para cada previsor deve estar distribuída entre as diferentes dimensões (ou autovalores). Para esse modelo, você poderá ver que cada previsor tem a maioria de suas variâncias em dimensões diferentes (a verba publicitária tem 96% da variância na dimensão 2, o número de execuções tem 93% da variância na dimensão 3 e a atratividade da banda tem 92% da variância na dimensão 4). Esses dados representam o exemplo clássico de multicolinearidade. Para um exemplo de quando a colinearidade existe nos dados e algumas sugestões sobre o que pode ser feito, veja os Capítulos 5 e 11 (Seção 6.8).

Dica da Samanta Ferrinho

Para checar as hipóteses de inexistência de multicolinearidade, use os valores FIV da tabela denominada **Coeficientes** (Coefficients) na saída do SPSS. Se esses valores forem menores do que 10, provavelmente não há motivo de preocupação. Se o valor médio da VIF e não for substancialmente maior do que 1, isso também indica que não há motivos para alarme.

Quadro 5.2

O que são autovalores e autovetores ④

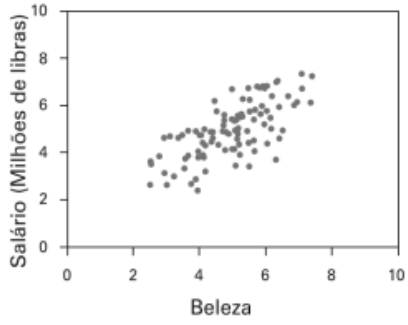
As definições e a matemática dos autovalores e autovetores são bastante complicadas e geralmente não precisamos nos preocupar com elas (embora eles surjam novamente nos Capítulos 14 e 15). Contudo, embora a matemática envolvida seja difícil, ela é fácil de visualizar! Imagine que temos duas variáveis: o salário anual de uma supermodelo e quão bela ela é. Imagine, ainda, que essas duas variáveis são normalmente distribuídas e, portanto, podem ser considerados em conjunto como uma distribuição normal bivariada. Se essas variáveis estiverem correlacionadas, o diagrama de dispersão forma uma elipse. Isso é mostrado nos diagramas de dispersão abaixo: se traçarmos uma linha pontilhada ao redor dos valores superiores do diagrama, obteremos algo de forma oval. Agora, podemos traçar duas linhas para medir o comprimento e a largura da elipse. Essas linhas são os autovetores da matriz de correlação original para essas variáveis (um vetor é apenas um conjunto de números que nos informa a localização de uma linha no espaço geométrico). Observe que as duas linhas traçadas (uma para a largura e outra para o comprimento da oval) são perpendiculares, isto é, elas formam um ângulo de 90 graus, o que significa que elas são independentes uma da outra. Assim, com duas variáveis, autovetores são apenas linhas medindo o comprimento e a largura da elipse ao redor do diagrama de dispersão dos dados para essas variáveis. Se adicionarmos uma terceira variável (por exemplo, a experiência da supermodelo), o diagrama de dispersão obtém uma terceira dimensão, a elipse torna-se algo com a forma de uma bola de rúgbi (ou de futebol americano), e porque adicionamos a terceira dimensão (comprimento, largura e altura), ganhamos um terceiro autovetor que mede a dimensão extra.

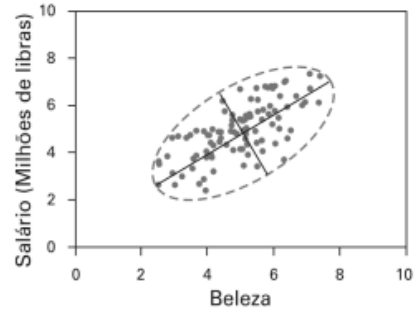
Se adicionarmos uma quarta variável, uma lógica semelhante é aplicada (embora seja mais difícil visualizá-la): obtemos uma dimensão extra e um autovetor para medir essa dimensão. Agora, cada autovetor tem um autovalor que nos informa seu comprimento (isto é, a distância de um ponto (início) a outro (final) do autovetor). Assim, examinando todos os autovetores de um conjunto de dados, sabemos as dimensões da elipse ou bola de rúgbi; ou seja, nós sabemos as dimensões dos dados. Portanto, os autovetores mostram quão equitativamente as variâncias da matriz estão distribuídas.

(*Continua*)

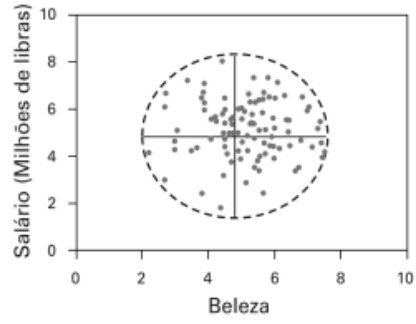
Material com direitos autorais

Quadro 5.2 (*Continuação*)





No caso de duas variáveis, a *condição* dos dados é relacionada à razão entre o tamanho do maior autovalor pelo do menor. Vamos olhar para os dois extremos: quando não existe relacionamento entre as variáveis e quando existe um relacionamento perfeito. Quando não existe relacionamento, o diagrama de dispersão irá estar mais ou menos contido em uma circunferência (ou em uma esfera, se tivermos três variáveis). Se novamente traçarmos linhas que medem o comprimento e a largura do círculo, veremos que elas apresentam o mesmo valor. Os autovalores medem o comprimento; portanto, os autovalores serão os mesmos. Assim, quando dividirmos o comprimento do maior autovalor pelo menor, obteremos 1 (porque os autovalores são os mesmos). Quando as variáveis são perfeitamente correlacionadas (isto é, existe uma colinearidade perfeita), o diagrama de dispersão forma uma linha reta e a elipse ao redor será muito estreita (com largura próxima de zero). Consequentemente, quando dividirmos o maior autovetor pelo menor, iremos obter um valor que tende para o infinito (porque o menor autovalor está próximo de zero.) Dessa forma, um índice de condição infinito é sinal de um grande problema.



5.8.6 Diagnósticos por casos (Casewise) ②

O SPSS produz uma tabela resumo de estatísticas residuais e elas devem ser examinadas para casos extremos. A Saída 5.10 do SPSS mostra casos com um resíduo padronizado menor que −2 ou maior do que 2 (lembre que mudamos o critério padrão de 3 para 2 na Figura 5.13). Mencionei na Seção 5.6.1.1 que em uma amostra comum esperamos que 95% dos casos tenham resíduos padronizados dentro de ±2. Nós temos uma amostra de 200, assim, é razoável esperar que aproximadamente 10 casos (5%) tenham resíduos padronizados fora desses limi-

Material com direitos autorais

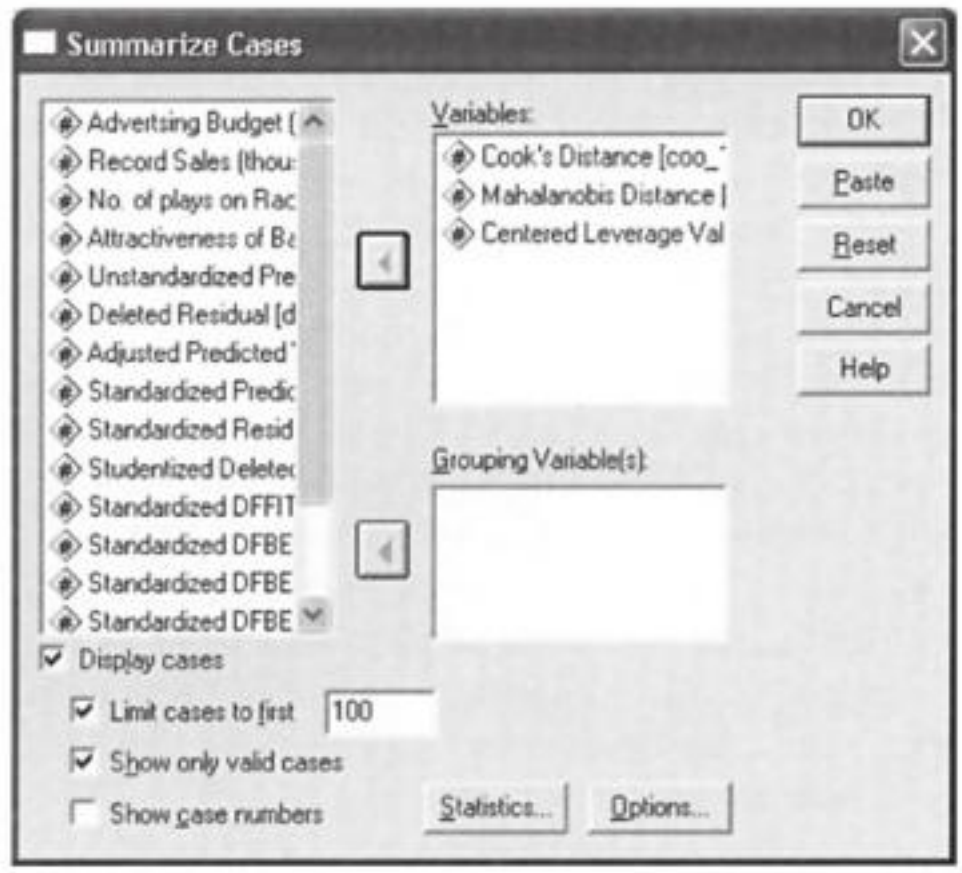


Figura 5.17

res duas (0,04) ou três vezes (0,06) maiores dependendo em qual estatístico você acredita mais (veja a Seção 5.6.1.2)! Todos os casos estão dentro dos limites de três vezes a média e somente o caso 1 está próximo de duas vezes a média. Finalmente, das recomendações para a distância de Mahalanobis, vimos que para uma amostra de 100 com três previsores, valores maiores do que 15 seriam problemáticos. Temos três previsores e uma grande amostra, assim, esse valor será um ponto de corte bem conservador, já que nenhum dos nossos casos está próximo de exceder esse critério. A evidência sugere que não existem

Saída 5.11 do SPSS

Case Summaries **(Resumo de Casos)**

	Case number (Caso número)	*Standardized DFBETA Intercept* (Intercepto do DFBETA padronizado)	*Standardized DFBETA ADVERTS* (PUBLICIDADE DFBETA padronizada)	*Standardized DFBETA AIRPLAY* (EXECUÇÕES DFBETA padronizadas)	*Standardized DFBETA ATTRACT* (ATRATIVIDADE DFBETA padronizada)	*Standardized DFFIT* (DFFIT padronizado)	*COVRATIO*
1	1	−0.31554	−0.24235	0.15774	0.35329	0.48929	0.97127
2	2	0.01259	−0.12637	0.00942	−0.01868	−0.21110	−0.92018
3	10	−0.01256	−0.15612	0.16772	0.00672	0.26896	0.94392
4	47	0.06645	0.19602	0.04829	−0.17857	−0.31469	0.91458
5	52	0.35291	0.02881	−0.13667	−0.26965	0.36742	0.95995
6	55	0.17427	0.32649	−0.02307	−0.12435	−0.40736	0.92486
7	61	0.00082	−0.01539	0.02793	0.02054	0.15562	0.93654
8	68	−0.00281	0.21146	−0.14766	−0.01760	−0.30216	0.92370
9	100	0.06113	0.14523	−0.29984	0.06766	0.35732	0.95888
10	164	0.17983	0.28988	−0.40088	−0.11706	−0.54029	0.92037
11	169	−0.16819	−0.25765	0.25739	0.16968	0.46132	0.85325
12	200	0.16633	−0.04639	0.14213	−0.25907	−0.31985	0.95435
Total N		12	12	12	12	12	12

(*Continua*)

***Case Summaries* (Resumo de Casos)**

	Case number (Caso número)	*Cook's Distance* (Distância de Cook)	*Mahalanobis Distance* (Distância de Mahalanobis)	*Centered Leverage Value* (Valor da Influência Centrado)
1	1	0.05870	8.39591	0.04219
2	2	0.01089	0.59830	0.00301
3	10	0.01776	2.07154	0.01041
4	47	0.02412	2.12475	0.01068
5	52	0.03316	4.81841	0.02421
6	55	0.04042	4.19960	0.02110
7	61	0.00595	0.06880	0.00035
8	68	0.02229	2.13106	0.01071
9	100	0.03136	4.53310	0.02278
10	164	0.07077	6.83538	0.03435
11	169	0.05087	3.14841	0.01582
12	200	0.02513	3.49043	0.01754
Total N		12	12	12

casos que podem influenciar nos nossos dados (embora todos os casos tenham que ser examinados para confirmar esse fato).

Podemos também examinar a estatística DFBeta para ver se algum caso tem uma grande influência nos parâmetros de regressão. Um valor absoluto maior do que 1 será um problema e em todos os casos os valores estão entre ± 1, o que mostra que esses casos não tem influência indevida sobre os parâmetros de regressão. Existe também uma coluna para a razão de covariâncias. Vimos na Seção 5.6.1.2 que precisamos utilizar o seguinte critério:

- $CVR_i > 1 + [3(k + 1)/n] = 1 + [3(3 + 1)/200] = 1{,}06$.
- $CVR_i < 1 - [3(k + 1)/n] = 1 - [3(3 + 1)/200] = 0{,}94$.

Dica da Samanta Ferrinho

Você precisa procurar casos que possam influenciar o modelo de regressão.

- Verifique os resíduos padronizados e confirme que não mais do que 5% dos casos apresentam valores absolutos acima de 2 e não mais do que 1% apresentam valores absolutos acima de 2,5. Qualquer caso com valor acima de 3 será um valor atípico.
- Olhe no editor de dados as distâncias de Cook: qualquer valor acima de 1 indica um caso que poderá influenciar o modelo.
- Calcule a influência média (o número de previsores mais 1 dividido pelo tamanho da amostra) e então observe valores maiores do que duas ou três vezes a influência média.
- Para a distância de Mahalanobis, uma verificação rápida é procurar valores acima de 25 em amostras grandes (500) e valores acima de 15 em amostras pequenas (100). Contudo, Barnett e Lewis (1978) devem ser consultados para uma análise mais detalhada.
- Procure valores com DFBeta maiores do que 1.
- Calcule os limites superior e inferior dos valores aceitáveis para a razão de covariâncias, CVR. O limite superior é 1 mais três vezes a influência média, enquanto que o limite inferior é 1 menos três vezes a influência média. Casos que apresentam um CVR fora desses limites podem ser problemáticos.

Material com direitos autorais

Portanto, estamos procurando por casos que se desviem substancialmente desses limites. Muitos dos nossos 12 valores atípicos potenciais apresentam valores CVR dentro ou próximo desses limites. O único caso que pode preocupar é o 169 (novamente!), cujo CVR está de alguma forma abaixo do limite inferior. Contudo, verificando a distância de Cook para esse caso, provavelmente não há motivo para preocupação.

Se você tivesse solicitado outras estatísticas de diagnóstico e com o que você já sabe de discussões anteriores sobre elas, você saberia que deve analisá-las na situação de casos não usuais nos dados. Contudo, a partir desse conjunto mínimo de diagnóstico parece que temos um modelo razoavelmente confiável que não tem sido influenciado por qualquer subconjunto de casos.

5.8.7 Conferindo as hipóteses ②

Para finalizar a análise, você deve verificar as hipóteses do modelo. Já checamos a colinearidade dentro dos dados e utilizamos a estatística de Durbin-Watson para verificar se os resíduos no modelo são independentes. Na Seção 5.7.3, solicitamos um gráfico dos *ZRESID contra *ZPRED e um histograma e um diagrama das probabilidades normais dos resíduos. O gráfico do *ZRESID e *ZPRED deve ser semelhante a um conjunto aleatório de pontos dispersos em torno de zero. Se esse gráfico afunila, é provável que os dados sejam heterocedásticos. Se existir qualquer tipo de curva nesse gráfico, as chances são de que esses dados não satisfazem a hipótese de linearidade. A Figura 5.18 mostra vários exemplos de gráficos dos resíduos padroniza-

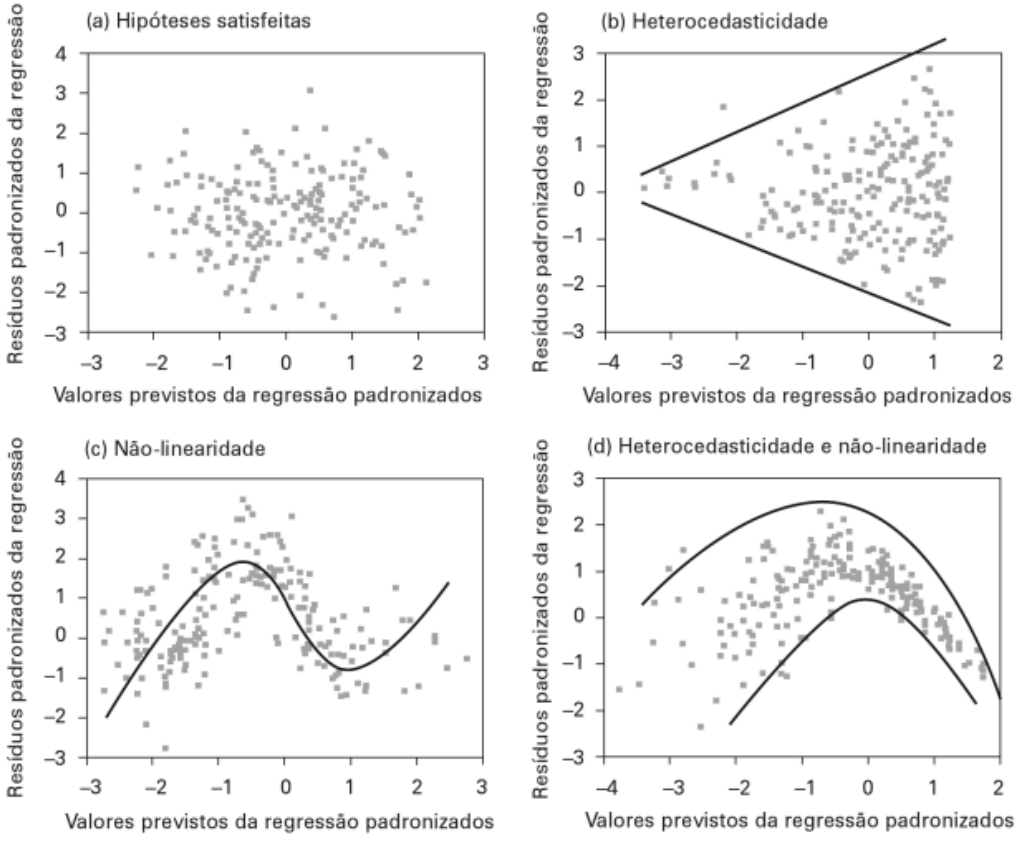


Figura 5.18 Diagramas dos *ZREDID contra *ZPRED.

Material com direitos autorais

dos contra os valores previstos padronizados. O painel (a) mostra o gráfico do nosso exemplo da venda de discos. Note como os pontos estão aleatoriamente dispersos por todo o quadrante. Esse padrão é um indicativo da situação em que as suposições de linearidade e homocedasticidade foram satisfeitas. O painel (b) mostra um diagrama semelhante para um conjunto de dados que viola a hipótese de homocedasticidade. Note que os pontos se espalham como uma forma de funil ficando cada vez mais distantes entre si à medida que os dados avançam para a direita. Essa é a forma típica da heterocedasticidade e indica um crescimento da variância nos resíduos. O painel (c) mostra um diagrama de dados que não tem um relacionamento linear entre os previsores e a variável de saída. Esse padrão é mostrado pelos resíduos. Uma linha ilustrando a relação não-linear foi desenhada sobre o gráfico destacando a tendência dos dados. Finalmente, o painel (d) representa uma situação na qual os dados não apenas apresentam um relacionamento não-linear, mas também heterocedasticidade. Note inicialmente que num canto do diagrama os pontos estão bem próximos, mas na outro eles estão bem separados. Quando essas suposições não forem satisfeitas você não irá ver exatamente esses padrões, mas espero que esses diagramas o ajude a entender os tipos de anomalias que você deve procurar.

Para testar a normalidade dos resíduos, devemos olhar para os histogramas e o diagrama de probabilidades normais selecionados na Figura 5.14. A Figura 5.19 mostra o histograma e o diagrama das probabilidades normais dos dados para o exemplo atual (lado esquerdo). O

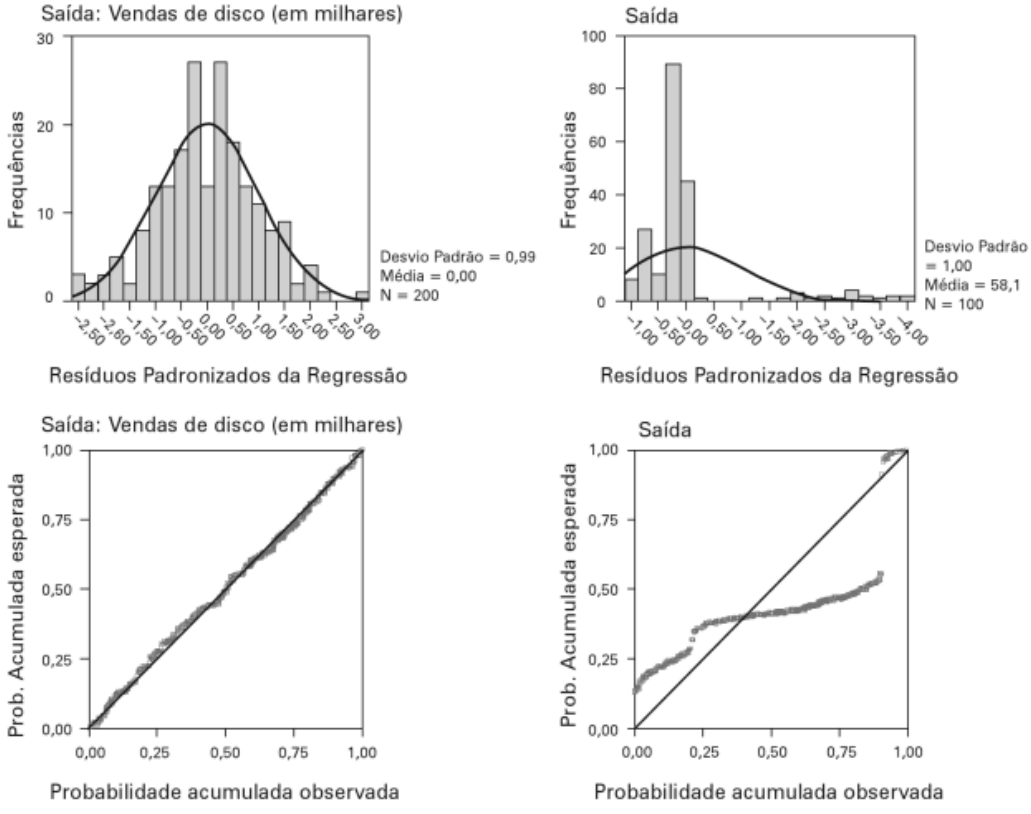
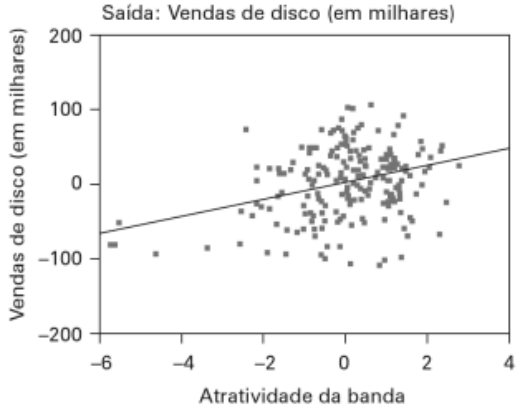


Figura 5.19 Histogramas e diagramas P-P dos resíduos normalmente distribuídos (lado esquerdo) e resíduos não normalmente distribuídos (lado direito).

Material com direitos autorais

histograma deve ser parecido ao de uma distribuição norma. O SPSS desenha a curva sobre o histograma para mostrar a forma da distribuição. Para os dados da gravadora, a distribuição é aproximadamente normal (embora exista uma leve deficiência dos resíduos exatamente em zero). Compare esse histograma com o que está ao lado dele, extremamente não-normal, e deve ficar claro que a distribuição não-normal apresenta uma assimetria acentuada. Assim, você deve procurar curvas que tenham a mesma forma que essa dos dados da venda de discos: qualquer desvio dessa curva é um sinal de não-normalidade e quanto maior a diferença, mais os resíduos estarão afastados da normalidade. O diagrama das probabilidades normais mostra também os desvios da normalidade (veja o Capítulo 2). A linha reta neste diagrama representa a distribuição normal e os pontos, os resíduos observados. Portanto, em um conjunto de dados perfeitamente normal; todos os pontos estarão bem próximos da linha. Isso é o que enxergamos para os dados da venda de discos. Contudo, próximo ao diagrama dos dados normalmente distribuídos da venda de discos está um exemplo de um afastamento extremo da normalidade. Neste diagrama, os pontos estão bem distantes da linha, o que indica um grande afastamento da normalidade. Para os dois diagramas, os dados não-normais são casos extremos e você deve ficar atento para desvios da normalidade mais sutis. É claro que você pode utilizar o que aprendeu no Capítulo 2 para realizar um teste K-S nos resíduos padronizados a fim de ver se eles se desviam ou não de forma significativa da normalidade.



Material com direitos autorais

Um conjunto final de diagramas, os diagramas parciais, foi especificado na Figura 5.14. Esses diagramas são dispersões dos resíduos da variável de saída e de cada um dos previsores quando as duas variáveis são modeladas separadamente dos demais previsores. Mencionei antes que valores atípicos óbvios em um gráfico parcial representam casos que podem ter influência nos coeficientes dos previsores da regressão e que relacionamentos não-lineares e heterocedásticos podem ser detectados, da mesma forma, utilizando esses diagramas.

Para os dados da verba publicitária, os diagramas parciais mostram um relacionamento positivo forte com as vendas de disco. O gradiente da linha é b para a verba publicitária no modelo (essa linha não é mostrada por omissão). Não existem valores atípicos óbvios nesse diagrama e a nuvem de pontos é igualmente espaçada em torno da linha, indicando homocedasticidade.

Para o número de execuções (*airplay*), os diagramas parciais mostram um relacionamento positivo forte com a venda de discos. O gradiente e o padrão dos dados são semelhantes aos da verba publicitária (o que já era esperado devido à semelhança dos beta padronizados desses previsores). Não existem valores atípicos óbvios nesse diagrama e a nuvem de pontos está igualmente espaçada em torno da linha indicando homocedasticidade.

Para a atratividade (*atractiveness*), os diagramas parciais mostram um relacionamento positivo forte com as vendas de discos. O relacionamento parece menos linear do que os demais previsores e os pontos parecem afunilados indicando um aumento da variância em níveis mais altos da atratividade. Não se percebem valores atípicos óbvios nesses dados, mas a nuvem de pontos em forma de funil pode indicar uma violação da hipótese de homocedasticidade. Seria adequado coletar alguns dados adicionais sobre bandas com baixa atratividade como forma de verificar o modelo atual.

Podemos resumir dizendo que o modelo parece, às vezes, em muitos sentidos, ser tanto preciso para a amostra e generalizável para a população. O único senão é se a atratividade da banda violou a hipótese de homocedasticidade. Portanto, podemos concluir que em nossa amostra a verba publicitária e o número de execuções são igualmente importantes em prever a venda de discos. A atratividade da banda é um previsor significativo da ven-

Dica da Samanta Ferrinho

Você precisa certificar-se de que algumas hipóteses da regressão foram satisfeitas a fim de ficar seguro de que ele pode ser generalizado para além da amostra.

- Verifique o gráfico do *ZRESID contra o *ZPRED. Se ele parecer com um conjunto aleatório de pontos, isso é bom. Se o gráfico parecer abrir como um funil de um lado ou outro, provavelmente teremos uma violação da hipótese de homogeneidade da variância. Se os pontos tiverem um padrão (forma curva), provavelmente teremos uma violação da linearidade. Se os dados parecem ter um padrão e estão mais espalhados em um ponto do que em outros, é provável que isso reflita uma violação tanto da homogeneidade da variância quanto da linearidade. Qualquer um desses cenários colocará a validade do nosso modelo em cheque. Repita a análise para os diagramas parciais também.
- Examine os histogramas e os diagramas P-P. Se os histogramas parecem com uma distribuição normal (e o diagrama P-P se assemelha a uma linha diagonal), tudo está bem. Se o histograma parece não-normal e o diagrama P-P se assemelha a uma cobra enrolada em torno da linha diagonal, as coisas não estão tão boas! Fique atento: distribuições normais podem parecer não-normais quando trabalhamos com pequenas amostras!

Material com direitos autorais

da de discos, mas é menos importante que os demais previsores (e provavelmente precisa de uma confirmação de uma possível heterocedasticidade). As hipóteses parecem ter sido satisfeitas e podemos provavelmente assumir que o modelo poderá ser generalizado para os novos discos sendo lançados.

5.9 COMO RELATAR A REGRESSÃO MÚLTIPLA ②

Se você seguir as recomendações da APA (American Psychological Association) para relatar a regressão múltipla, os resultados tabelados devem ser como os apresentados adiante. A APA também exige que sejam apresentados os betas padronizados, seus valores da significância e algumas estatísticas gerais sobre o modelo (tal como o R^2). Se você decidir fazer uma tabela, os valores beta e seus erros padrão também serão bastante úteis assim como a constante, para que os leitores do seu trabalho possam reproduzir o modelo completo se necessário. Ainda, se você executou uma regressão hierárquica, deve relatar esses valores em cada estágio da hierarquia. Desse modo, basicamente, você vai reproduzir a tabela denominada de coeficientes (*coefficients*) da saída do SPSS omitindo as informações não-essenciais. Para o exemplo deste capítulo podemos produzir uma tabela como a 5.2.

Assim, você poderá retomar a algumas das saídas do SPSS neste capítulo e descobrir de onde esses valores vieram. As coisas a serem destacadas são: (1) arredondamos todos os valores para a segunda casa decimal; (2) para os betas padronizados não existe o zero antes da vírgula (porque esses valores não podem exceder 1),* mas para todos os demais valores menores do que um o zero está presente; (3) a significância de uma variável é representada por um asterisco com uma nota de rodapé a fim de indicar o nível de significância sendo utilizado (se existir mais do que um nível de significância, você pode representar isso com múltiplos asteriscos, tal como, $*p < 0{,}05$, $**p < 0{,}01$ e $***p < 0{,}001$) e (4) o R^2 para o modelo inicial e a mudança no R^2 (representada por ΔR^2) para cada passo subsequente do modelo são apresentados embaixo da tabela.

Tabela 5.2 Como relatar a regressão múltipla

	B	*SE B*	*β*
Passo 1			
Constante	134,14	7,54	
Verba publicitária	0,10	0,01	0,58*
Passo 2			
Constante	– 26,61	17,35	
Verba publicitária	0,09	0,01	0,51*
Execuções na Radio 1 da BBC	3,37	0,28	0,51*
Atratividade	11,09	2,44	0,19*

Nota: $R^2 = 0{,}34$ para o Passo 1; $\Delta R^2 = 0{,}33$ para o passo 2 ($ps < 0{,}001$). * $p < 0{,}001$.

5.10 PREVISORES CATEGÓRICOS E A REGRESSÃO MÚLTIPLA ③

Muitas vezes na análise de regressão, coletamos dados sobre grupos de pessoas (alguns exemplos das ciências sociais podem ser o grupo étnico, o gênero, o status socioeconômico e a categoria de diagnóstico). Você pode querer esses grupos como previsores no modelo de regressão; contudo, vimos que entre as hipóteses do modelo está a necessidade de as variáveis serem contínuas ou pelo menos categóricas com apenas duas categorias. Vimos na Seção 4.5.6 que a correlação ponto-bisserial é o *r* de Pearson entre duas variáveis quando uma é contínua e a outra tem duas categorias codificadas como 0 e 1. Também aprendemos que a regressão simples é baseada no *r* de Pearson, assim, não será necessário muita imaginação para ver que, como a correlação ponto-bisserial, podemos construir um modelo de regressão com um previsor que tem duas categoriais (por exemplo, gênero). Da mesma forma, não deve ser tão inconcebível perceber que podemos estender esse modelo para incorporar vários previsores que tem duas

* N. de T.: Essa observação não cabe aqui, pois os valores no Brasil sempre são escritos com o zero antes da vírgula, diferentemente do sistema anglo-saxão que costuma representar números entre −1 e 1 sem o zero antes do ponto decimal (vírgula para nós).

Material com direitos autorais

riais foi rotulada de "ecléticos". No arquivo de dados, ela codificou esses grupos como 1, 2, 3 e 4, respectivamente.

A primeira coisa que você deve fazer é calcular o número de variáveis *dummy*. Se temos quatro grupos, deverão existir três variáveis *dummy* (uma a menos que o número de grupos ou categorias). Depois precisamos definir o grupo-base. Queremos comparar as pessoas que têm afiliação musical com as que não têm, assim, a nossa categoria base será "ecléticos". Daremos a esse grupo o código 0 para todas as nossas variáveis *dummy*. Para a primeira variável *dummy,* podemos olhar para grupo *punk* e para fazer isso daremos a todos os *punks* o código 1, e aos demais o código 0. Para a nossa segunda variável *dummy*, usaremos o grupo dos "metaleiros", e daremos a qualquer pessoa que pertença a esse grupo o código 1, e aos outros o 0. Temos ainda uma variável *dummy* sobrando, destinada ao grupo "pagodeiros": vamos dar a cada pagodeiros o código 1 e aos demais, o código 0. O esquema final é mostrado na Tabela 5.3. Você deve observar que em cada grupo o código 1 foi utilizado em apenas uma variável *dummy* (exceto na categoria base que é sempre codificada como 0).

Como já disse, iremos prestar atenção em como essa codificação funciona na Seção 7.8 e 8.2.2; por enquanto, verifique como codificar nossa variável de agrupamento nessas variáveis *dummy* utilizando o SPSS. Para codificar variáveis, você precisa utilizar a função *Recode* (recodificar), Use **Transform⇒Recode⇒Into Different Variables**... (Transformar⇒Recodificar⇒Em diferentes variáveis) para acessar a caixa de diálogo vista na Figura 5.20. A primeira caixa de diálogo lista

Tabela 5.3 Codificação *dummy* para os dados do festival de Glastonbury

	Variável *dummy* 1	Variável *dummy* 2	Variável *dummy* 3
Ecléticos	0	0	0
Pagodeiros	0	0	1
Metaleiros	0	1	0
Punks	1	0	0

todas as variáveis no editor de dados, e você precisa selecionar as que quiser recodificar (nesse caso, **music** – música) e transferi-las para o painel denominado *Numeric Variable → Output Variable* (Variável numérica → Variável de Saída) clicando em ▸. Você precisa então dar um nome para a nova variável (a variável de saída, de acordo com a denominação do SPSS). Para tanto, vá para a parte denominada *Output Variable* (Variável de Saída) e no espaço onde diz *Name* (Nome), escreva o nome para a sua primeira variável *dummy* (você poderá chamá-la de **music1**). Você deve dar a essa variável um nome mais descritivo digitando algo no espaço denominado *Label* (Rótulo) (denominamos essa primeira variável *dummy* de ecléticos × *punks*). Quando você tiver feito isso, clique em Change a fim de transferir essa nova variável para o painel denominado *Numeric Variable → Output Variable* (Variável numérica → Variável de Saída); esse painel deve apresentar agora a seguinte legenda *music → music1*.

Uma vez definida a primeira variável *dummy*, precisamos informar ao SPSS como recodificar os valores da variável **music** nos novos valores que queremos para a nova variável **music1**. Para tanto, clique em Old and New Values... a fim de abrir a segunda caixa de diálogo como a da Figura 5.20. Essa caixa de diálogo é utilizada a fim de mudar os valores da variável original para valores diferentes da nova variável. Para a primeira variável *dummy*, queremos que qualquer um que seja *punk* tenha um código de 1 e os demais, um código de 0. *Punks* foram codificados com o valor 3 na variável original, assim, é preciso digitar 3 na seção denominada *Old Value* (Valor Antigo) no espaço chamado de *Value* (Valor). O novo valor que queremos é 1, dessa forma, será necessário digitar o valor 1 na seção denominada *New Value* (Valor novo) no espaço chamado de *Value* (Valor). Depois, clique em Add para adicionar isso a lista de alterações (a lista é mostrada no painel denominado *Old → New* (Velho → Novo), em que deve agora deve estar escrito 3 → 1, como no diagrama). O próximo passo é alterar os grupos remanescentes que apresen-

Material com direitos autorais

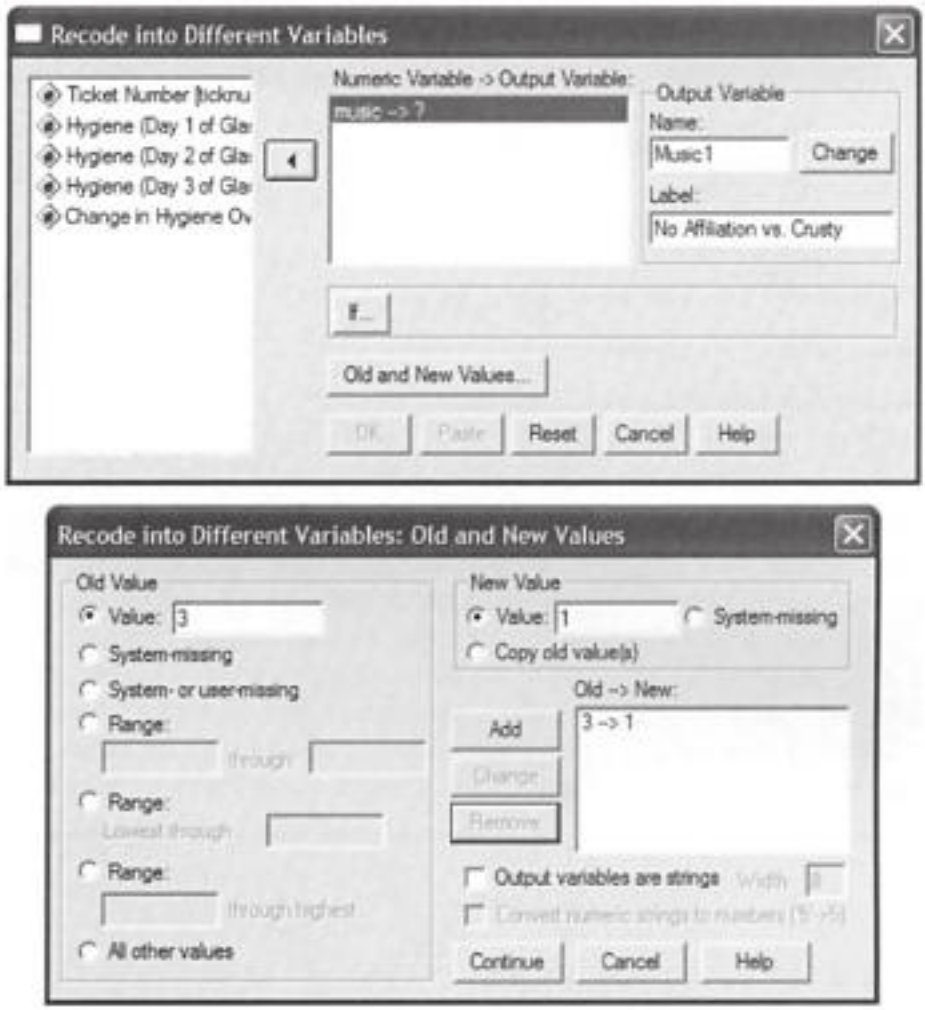


Figura 5.20 Caixa de diálogo para a função *Recode* (recodificar).

tam o valor 0 para a primeira variável *dummy*. Para tanto, selecione All other values e digite o valor 0 na seção denominada *New Value* (Valor Novo) no espaço *Value* (Valor). Quando tiver feito isso, clique em Add a fim de adicionar essa alteração à lista de mudanças (essa lista está mostrando *ELSE* → *0*). Depois, clique em Continue para retornar à caixa de diálogo principal e clique em OK a fim de criar a primeira variável *dummy*. Essa variável irá aparecer agora em uma nova coluna no editor de dados e você irá notar que ela tem um valor de 1 para qualquer pessoa originalmente classificada como *punk* e o valor 0 para os demais. Agora, tente criar as próximas duas variáveis *dummy* (denomine-as de **Music2** e **Music3**) utilizando os mesmos princípios.

5.10.2 Saída do SPSS para variáveis auxiliares (*dummy*) ③

Vamos assumir que criamos as variáveis de código ou auxiliares (se você ainda não conseguiu, use o arquivo de dados denominado **GlastonburyDummy.sav**). Você precisa entrar com todas as variáveis auxiliares relacionadas no mesmo bloco (assim, utilize o método *Enter*). Neste caso, teremos que entrar com as variáveis *dummy* no mesmo bloco; contudo, se tivéssemos outra variável (por exemplo, status socioeconômico) que tivesse sido transformada em uma variável *dummy*, poderíamos inseri-la em um bloco diferente (assim, apenas as variáveis *dummy* recodificadas de uma mesma variável precisam ser

Material com direitos autorais

Quadro 5.3

Utilizando a sintaxe para recodificar ③

Se você está fazendo muitas recodificações, logo irá cansar de utilizar a caixa de diálogo todas as vezes. No *site* www.artmed.com.br, há um arquivo de sintaxe, **RecodeGlastonburyData.sps**, que eu escrevi para criar todas as variáveis *dummy* que foram discutidas. Carregue esse arquivo e rode a sintaxe, ou abra a janela da sintaxe (veja a Seção 2.6) e digite o seguinte:

```
RECODE music (3 = 1) (ELSE = 0) INTO Music1.
RECODE music (2 = 1) (ELSE = 0) INTO Music2.
RECODE music (1 = 1) (ELSE = 0) INTO Music3.
VARIABLE LABELS Music1 'No affiliation  vs. Crusty'. Ecleticos vs. Punks
VARIABLE LABELS Music2 'No affiliation vs. Metaller'. Ecleticos vs. Metaleiros
VARIABLE LABELS Music3 'No affiliation vs. Indies Kid'. Ecleticos vs. Pagodeiros
EXECUTE.
```

Cada comando RECODE realiza o equivalente ao que eu fiz quando utilizei a caixa de diálogo apresentada na Figura 5.20. Assim, as primeiras três linhas apenas criam três novas variáveis (**Music1**, **Music2** e **Music3**), baseadas na variável original **Music**. Para essa primeira variável, se **music** é igual a 3, então ela torna-se 1, e todos os demais valores tornam-se 0. Para a segunda, se **music** é igual a 2, ela torna-se 1 e os demais valores ficam iguais a 0. A mesma lógica vale para a terceira variável *dummy*. As linhas iniciando com VARABLES LABELS dizem ao SPSS para atribuir o texto entre apóstrofes como rótulos para as variáveis **Music1**, **Music2** e **Music3**, respectivamente. A linha final tem o comando EXECUTE, sem o qual nenhum dos comandos anteriores será executado. Observe ainda que cada linha termina com um ponto final.

colocadas em blocos). Use o que você aprendeu neste capítulo para rodar uma regressão múltipla utilizando os escores alterados como saída e as três variáveis auxiliares (colocadas em blocos) como previsores. Vamos dar uma olhada na saída.

A Saída 5.12 do SPSS mostra as estatísticas do modelo. Isso mostra que entrando com as três variáveis *dummy*, podemos explicar 7,6% da variância dos escores da higiene (o valor R^2). Em outras palavras, 7,6% da variância nos escores da higiene pode ser explicada pela afiliação musical da pessoa. A ANOVA (que mostra o mesmo que a estatística de mudança R^2 porque existe apenas um passo nessa regressão) nos informa que o modelo é significantemente melhor em prever a mudança nos escores da higiene do que não ter um modelo (ou seja, 7,6% da variância explicada é uma quantia significativa). Isso deve ficar claro com base no que você leu neste capítulo; o que é mais interessante é como interpretar as variáveis *dummy* individualmente.

A Saída 5.13 do SPSS mostra a tabela básica dos coeficientes para as variáveis *dummy* (excluí os intervalos de confiança e o diagnóstico de colinearidade). A primeira coisa a observar é que cada variável *dummy* aparece na tabela com um rótulo útil (tal como *punks versus* ecléticos). Isso acontece porque quando codificamos as nossas variáveis, fornecemos um rótulo a cada uma; se não tivéssemos feito isso, a tabela conteria os nomes bem menos úteis das variáveis (music1, music2 e music3). Os rótulos que eu sugeri que fossem fornecidos para as variáveis nos dão uma boa dica sobre o que cada variável representa. A primeira variável *dummy* (ecléticos *versus punks*) mostra a diferença entre a mudança nos escores da higiene para os grupos ecléticos e *punks*.

Material com direitos autorais

Saída 5.12 do SPSS

***Model Summary*[b] (Resumo do Modelo)**

Model (Modelo)	*R*	*R Square* (R quadrado)	*Adjusted R Square* (R Quadrado Ajustado)	*Std. Error of the Estimate* (Erro Padrão da Estimativa)	*Change Statistics* (Estatísticas de Mudança)					*Durbin-Watson*
					R Square Change (Mudança do R quadrado)	*F Change* (Mudança no F)	*df1* (gl1)	*df2* (gl2)	*Sig. F Change* (Sig. da Mudança no F)	
1	0.276[a]	0.076	0.053	0.68818	0.076	3.270	3	119	0.024	1.893

a. Previsores (Constante) Ecléticos *vs*. Pagodeiros, Ecléticos *vs*. *Punks*, Ecléticos *vs*. Metaleiros.
b. Variável Dependente: Mudança na higiene ao longo do festival.

***ANOVA*[b] (Análise de Variância)**

Model (Modelo)		*Sum of Squares* (Soma dos quadrados)	*df* (gl)	*Mean Square* (Média dos Quadrados)	*F*	*Sig.*
1	*Regression* (Regressão)	4.646	3	1.549	3.270	0.024[a]
	Residual (Resíduos)	56.358	119	0.474		
	Total	61.004	122			

a. Previsores (Constante) Sem Afiliação *vs*. pagodeiros, Sem afiliação *vs*. *punks*, Sem Afiliação *vs*. Metaleiros.
b. Variável Dependente: Mudança na higiene ao longo do festival.

Lembre-se de que os valores beta nos informam a mudança na saída devido a uma unidade no previsor. Nesse caso, a mudança de uma unidade no previsor é uma alteração de 0 para 1. Ela mostra a mudança nos escores da higiene como resultado mudança da variável *dummy* de 0 para 1 (*punk*). Pela inclusão das três variáveis ao mesmo tempo, nossa categoria base é sempre 0; assim, isso representa a mudança nos escores da higiene se uma pessoa eclética é comparada com um *punk*. Essa mudança é a diferença entre as médias dos dois grupos. Para ilustrar isso, produzi uma tabela com as médias de cada um dos quatro grupos. Essas médias representam a mudança média nos escores da higiene para os três grupos (isto é, os homens de cada grupo na nossa variável de saída). Se calcularmos as diferenças entre essas médias para o grupo dos ecléticos e o dos *punks*, obteremos: *Punks* – Ecléticos = (−0,9658) − (−0,5543) = −0,4120. Ou seja, a mudança nos escores da higiene é maior para o grupo dos *punks* do que para os ecléticos (a higiene dos *punks* decresce mais ao longo dos três dias do festival do que a dos ecléticos). Esse valor é o mesmo que o beta na Saída 5.13 do SPSS! Assim, os valores beta nos informam a diferença relativa entre cada grupo e o grupo que foi escolhido como categoria base. Esse valor beta é convertido em uma estatística *t* e a significância dessa estatística é calculada. Essa estatística está testando, como já foi visto antes, se o valor é 0, e quando nós temos duas categorias codificadas com 0 e 1, isso significa que o que está sendo testado é se a diferença entre grupos é zero. Se ela é significativa, isso quer dizer que o

***Report* (Relatório)**

Change in Hygiene Over The Festival
(Mudança da Higiene ao Longo do Festival)

Musical Affiliation (Afiliação Musical)	*Mean* (Médias)	*N*	*Std Deviation* (Desvios Padrão)
Indie Kid (Pagodeiros)	– 0.9643	14	0.67020
Metaller (Metaleiros)	– 0.5259	27	0.57583
Crusty (*Punks*)	– 0.9658	24	0.76038
No Musical affiliation (Ecléticos)	– 0.5543	58	0.70834
Total	– 0.6750	123	0.70713

Material com direitos autorais

Saída 5.13 do SPSS

Coefficients[a] (Coeficientes)					
	Unstandardized Coefficients (Coeficientes não-padronizados)		*Standardized Coefficients* (Coeficientes padronizados)		
Model (Modelo)	*B*	*Std. Error* (Erro Padrão)	*Beta*	*t*	*Sig.*
1 *(Constant)* (Constante)	– 0.554	0.090		– 6.134	0.000
No affiliation vs. Crusty (Ecléticos *vs. Punks*)	–0.412	0.167	– 0.232	– 2.464	0.015
No affiliation vs. Metaller (Ecléticos *vs.* Metaleiros)	0.028	0.160	0.017	0.177	0.860
No affiliation vs. Indie Kid (Ecléticos *vs.* Pagodeiros)	– 0.410	0.205	– 0.185	– 2.001	0.048

a. Variável dependente: mudança na higiene ao longo do festival.

grupo codificado como 1 é significativamente diferente da categoria base – assim, ela está testando a diferença entre duas médias, que é o contexto onde você estará mais familiarizado com a estatística *t* (veja o Capítulo 7). Para a nossa primeira variável *dummy*, o teste *t* é significativo, e o valor beta é negativo; assim, podemos dizer que a mudança na higiene decresce quando uma pessoa passa do grupo dos ecléticos para o grupo dos *punks*. Tenha em mente que um decréscimo nos escores da higiene representa uma mudança maior (você está ficando mais fedorento), ou seja, a higiene decresce significativamente mais nos *punks* do que nos ecléticos.

A próxima variável *dummy* compara metaleiros a ecléticos. O valor beta novamente representa a mudança na higiene se uma pessoa eclética é comparada com um metaleiro. Se calcularmos a diferença entre as médias dos grupos para o grupo eclético e o dos metaleiros, obteremos: metaleiro – eclético = (-0,5259 – (−0,5543) = 0,028. Esse valor é novamente igual ao valor beta na Saída 5.13 do SPSS! Para essa segunda variável *dummy*, o teste *t* não é significativo, assim, podemos dizer que a mudança dos escores da higiene é o mesmo se uma pessoa muda do grupo eclético para o metaleiro. Em outras palavras, a mudança nos escores da higiene não é previsível se uma pessoa é metaleira comparada com uma eclética.

Em nossa última variável *dummy*, comparamos *punks* com ecléticos. O valor beta novamente representa a alteração nos escores da higiene se uma pessoa sem afiliação musical é comparada com um pagodeiro. Se calcularmos a diferença nas médias dos grupos entre os sem afiliação e os pagodeiro, obteremos: pagodeiro – sem afiliação = (−0,9643) − (−0,5543) = −0,410. Não deve ser mais surpresa agora que esse valor é o beta da Saída 5.13 do SPSS! O teste *t* é significativo e o beta tem um valor negativo, assim, como ocorreu com a primeira variável *dummy*, podemos dizer que a mudança nos escores da higiene diminui se uma pessoa muda de sem afiliação para pagodeiro. Lembre-se de que um decréscimo nos escores da higiene representa mais alterações (você está se tornando mais fedido); ou seja, pagodeiros são significativamente mais fedorentos do que os ecléticos!

De forma geral, essa análise mostrou que, comparado com o grupo ecléticos, os grupos de *punks* e pagodeiros ficam mais fedidos ao longo dos três dias do festival, mas o grupo de metaleiros não. Esta seção introduziu algumas ideias complexas que irei detalhar nos Capítulos 7 e 8. Se você estiver muito confuso ou quiser saber mais sobre como variáveis *dummy* funcionam, sugiro que você leia as

Material com direitos autorais

- Regressão múltipla
- Regressão passo a passo
- Regressão simples
- Resíduo
- Resíduo excluído
- Resíduos estudentizados
- Resíduos excluídos estudentizados
- Resíduos não-padronizados
- Resíduos padronizados
- Soma dos quadrados do modelo
- Soma dos quadrados dos resíduos
- Soma total dos quadrados
- Teste de Durbin-Watson
- Tolerância
- Validação cruzada
- Valor previsto ajustado
- Valores Chapéu
- Variável de saída
- Variável *dummy*
- Variável previsora

5.13 TAREFAS DO ALEX ESPERTO

As respostas para essas tarefas podem ser encontradas no *site* www.artmed.com.br no arquivo **(Chapter 5).pdf.**

- Uma estudante de moda está interessada em fatores que possam prever os salários de modelos de passarela. Ela coletou dados de 231 modelos. Para cada modelo, ela perguntou o salário diário nos dias em que elas estavam trabalhando (**salary**), a idade (**age**), quanto tempo elas estão trabalhando como modelos (**years**). A estudante também reuniu um painel de especialistas de agências de modelo a fim de dar uma nota para a atratividade de cada modelo, com uma percentagem de 100% representando o máximo de atratividade (**beauty**). Os dados estão no *site* www.artmed.com.br no arquivo **Supermodel.sav**. Infelizmente, essa estudante de moda comprou alguns textos de estatística não muito bons e não sabe como analisar os seus dados. ☺ Você pode ajudá-la conduzindo uma regressão múltipla para verificar quais fatores preveem o salário da modelo? Quão válido é o modelo de regressão? ②
- Utilizando os dados de Glastonbury deste capítulo (com os códigos *dummy* em **GlastonburyDummy.sav**), que você já analisou, comente se você julga que o modelo é confiável e generalizável. ③

5.14 LEITURAS COMPLEMENTARES

BOWERMAN, B. L, O'CONNELL, R. T. *Linear statistical models: an applied approach*. Belmont (CA): Duxbury, 1990. 2ª ed. Esse texto é destinado somente para os que gostam de matemática ou estudantes de pós-graduação, mas fornece uma exposição extremamente abrangente da análise de regressão.

HARDY, M. A. *Regression with dummy variables*. Sage university paper series on quantitative applications in the Social Sciences. 07-093. Newbury Park (CA): Sage, 1993.

HOWEL, D. C. *Statistical methods for psychology*. Belmont (CA): Duxbury, 2002. 5ª ed. Os Capítulos 9 e 15 são excelentes introduções à matemática por trás dos modelos de regressão.

MILES, J., SHEVILIN, M. *Applying regression and correlation: a guide for students and researchers*. London: Sage, 2001. Esse texto bastante simples aborda a regressão em detalhes com um mínimo de esforço. Altamente recomendável.

STEVENS, J. *Applied multivariate statistics for the social sciences*. Hillsdale (NJ): Erlbaum, 1992. 2ª ed. Capítulo 3.

Material com direitos autorais

6
REGRESSÃO LOGÍSTICA

6.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ①

No capítulo anterior, vimos como modelar o relacionamento entre uma ou mais variáveis previsoras e uma saída. Este capítulo amplia esse assunto para verificar como podemos prever uma variável de saída categórica utilizando a *regressão logística*. É um tema difícil (de fato, eu poderia tê-lo enchido de quadros do Alex Esperto, mas não o fiz!) e a regressão logística não é um dos meus pontos fortes, assim, provavelmente escrevi algumas bobagens. No entanto, começamos com um pouco de teoria para ajudá-lo a entender a regressão logística, antes de irmos direto para um exemplo. Você irá aprender a executar a regressão logística no SPSS e a interpretar a saída resultante. Concluímos o capítulo com um segundo exemplo e usaremos esse exemplo a fim de ver como combater a multicolinearidade. Ao longo do texto, também descobriremos por que a seleção de futebol inglesa não sabe cobrar pênaltis.

6.2 PRESSUPOSTOS DA REGRESSÃO LOGÍSTICA ①

Em poucas palavras, a regressão logística é uma regressão múltipla, mas com uma variável de saída categórica dicotômica e variáveis previsoras contínuas ou categóricas. Simplificando, isso quer dizer que podemos prever a qual de duas categoriais é provável que uma pessoa pertença dado certas informações. Um exemplo trivial é determinar que variáveis podem definir se uma pessoa é homem ou mulher. Você poderá mensurar a preguiça, a estupidez, o consumo de álcool, o número de arrotos que a pessoa dá em um dia. Utilizando a regressão logística, podemos verificar que todas essas variáveis são capazes de prever o gênero de uma pessoa, e a técnica nos apontará também se uma pessoa tem certa probabilidade de ser homem ou mulher. Assim, se você pegar uma pessoa ao acaso e descobrir que ela teve um escore alto em preguiça, estupidez, consumo de álcool e número de arrotos, então o modelo de regressão poderá nos dizer, com base nessa informação, quão provável é que essa pessoa seja homem.

É improvável que um pesquisador esteja interessado no relacionamento entre flatulência e gênero (isso é suficientemente comprovado pela experiência para justificar uma pesquisa!), mas a regressão logística pode ter aplicações vitais. Na literatura biomédica (isto é, pesquisa médica), a regressão logística tem aplicações tais como a de formular modelos sobre os tipos

Material com direitos autorais

de fatores que determinam se um tumor é cancerígeno ou benigno. Uma base de dados de pacientes pode ser utilizada para identificar as variáveis que são influentes na previsão do tipo de um tumor. Essas variáveis podem então ser mensuradas em um novo paciente e seus valores colocados no modelo de regressão logística a partir do qual é possível estimar uma probabilidade de o tumor ser maligno. Se a probabilidade do tumor ser maligno é baixa, o médico pode decidir não executar uma cara e dolorosa cirurgia que seria desnecessária.

Nas ciências sociais, raramente enfrentamos decisões de vida ou morte, mas mesmo assim a regressão logística é uma ferramenta bastante útil. Por esse motivo, é triste que muitos livros-texto abordem o tema superficialmente. Neste capítulo, explicarei os princípios por trás da regressão logística e como executar os procedimentos no SPSS.

6.3 QUAIS SÃO OS PRINCÍPIOS POR TRÁS DA REGRESSÃO LOGÍSTICA? ③

Não pretendo entrar nos princípios subjacentes da regressão logística porque eles não são necessários para entender o teste (eu sou uma prova viva disso). Contudo, quero fazer algumas comparações com a regressão usual a fim de inserir a regressão logística em uma estrutura que seja familiar a qualquer um que tenha chegado até aqui no livro. (O quê? Você não leu o capítulo de regressão ainda?!) Se você tem fobia de equações, é melhor olhar para o outro lado. Em uma regressão linear simples, vimos que a variável de saída Y é prevista a partir da equação da linha:

$$Y = b_0 + b_1X + \varepsilon \qquad (6.1)$$

Onde b_0 é o intercepto e b_1 é o gradiente da linha, X é o valor da variável previsora e ε é o termo resíduo. Dados os valores de Y e X, os parâmetros desconhecidos na equação podem ser estimados encontrando-se a solução para a qual as distâncias ao quadrado entre os valores observados e previstos da variável dependente sejam mínimas (método dos mínimos quadrados).

Esse assunto deve ser bem familiar. Na regressão múltipla – em que existem vários previsores, – uma equação semelhante é derivada na qual cada previsor tem seu próprio coeficiente. Como tal, Y é previsto a partir de uma combinação de cada variável previsora multiplicada pelo seu respectivo coeficiente de regressão:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon \qquad (6.2)$$

Onde b_n é o coeficiente de regressão da correspondente variável X_n. Na regressão logística, em vez de se prever o valor da variável Y a partir de um previsor X ou diversas variáveis previsoras (Xs), prevemos a *probabilidade* de Y ocorrer conhecidos os valores de X ou Xs. A equação de regressão logística apresenta várias semelhanças com a equação de regressão recém-descrita. Na sua forma mais simples, quando existe um único previsor X, a equação de regressão logística a partir da qual a probabilidade da variável Y é prevista é dada por:

$$P(Y) = \frac{1}{1 + e - (b_0 + b_1X + \varepsilon_i)} \qquad (6.3)$$

Onde $P(Y)$ é a probabilidade de Y ocorrer, e é a base dos logaritmos naturais e os demais coeficientes da equação formam uma combinação linear muito semelhante à regressão simples. De fato, você pode ter notado que a parte da equação dentro dos parênteses é idêntica à regressão linear, pois existe uma constante (b_0), uma variável previsora (X) e um coeficiente (ou peso) agregado ao previsor (b_1). Da mesma forma que para a regressão linear, é possível estender essa equação para incluir diversas variáveis previsoras. Quando isso ocorre, a equação toma a seguinte forma:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon)}} \qquad (6.4)$$

A equação (6.4) é a mesma utilizada quando existe um único previsor, exceto que a combinação linear foi estendida para incluir qualquer número de previsores. Assim, enquanto que a versão com um previsor da equação de regressão logística contém a equação

Material com direitos autorais

de regressão linear simples em seu interior, a versão com múltiplos previsores contém a equação de regressão múltipla.

A despeito das semelhanças entre a regressão linear e a logística, existe uma boa razão porque não podemos aplicar a regressão linear diretamente a uma situação onde a variável de saída é dicotômica: uma das hipóteses da regressão linear é que o relacionamento entre as variáveis é linear (veja a Seção 5.6.2.1). Vimos anteriormente a importância das hipóteses para se obter um modelo preciso. Portanto, para o modelo de regressão linear ser válido, os dados observados devem ter um relacionamento linear. Quando a variável de saída é dicotômica, essa hipótese é normalmente violada (veja Berry, 1993). Uma forma de contornar esse problema é alterar os dados por meio de uma transformação logarítmica (veja Berry e Feldman, 1985, Capítulo 3). Isso mantém a *forma* do relacionamento linear enquanto o próprio relacionamento em si é não-linear (assim, é um modo de expressar um relacionamento não-linear em uma maneira linear). A equação de regressão logística descrita anteriormente está baseada no seguinte princípio: ela expressa uma equação de regressão linear múltipla em termos logarítmicos e dessa forma resolve o problema da violação da hipótese de linearidade.

A forma exata da equação pode ser arranjada de várias maneiras, mas a versão que eu escolhi expressa a equação em termos da probabilidade de Y ocorrer (isto é, a probabilidade de que um caso pertença a uma determinada categoria). O valor resultante da equação é a uma probabilidade e, como tal, varia sempre entre 0 e 1. Um valor próximo de 0 significa que a ocorrência de Y é bastante improvável e um valor próximo de 1, que ela é bem provável. Além disso, da mesma forma que na regressão linear, na equação logística cada variável previsora tem seu próprio coeficiente. Quando executamos a análise precisamos estimar os valores desses coeficientes para que possamos utilizar a equação. Esses parâmetros são estimados pelo ajustamento de modelos, com base nos previsores disponíveis, aos dados observados. O modelo escolhido será aquele que quando os valores das variáveis previsoras forem utilizados resulta em um valor de Y mais próximo do valor observado. Especificamente, os valores dos parâmetros são estimados utilizando a **estimação de máxima verossimilhança** que seleciona os coeficientes que tornam os valores observados mais prováveis de terem ocorrido. Assim como ocorreu com a regressão múltipla, tentamos ajustar um modelo aos nossos dados que nos permita estimar valores da variável de saída a partir de valores das variáveis previsoras.

6.3.1 Avaliando o modelo: a estatística de verossimilhança-log ③

Como vimos, o modelo de regressão logística prevê a probabilidade de um evento ocorrer para uma dada pessoa (representaríamos isso como $P(Y_i)$, a probabilidade de que Y ocorra para a pessoa i), baseado em observações de se um evento ocorreu ou não para essa pessoa (representaríamos isso como Y_i,, a saída real para a i-ésima pessoa). Assim, para uma dada pessoa, Y poderá ser 0 (a saída não ocorre) ou 1 (a saída ocorre), e o valor previsto, $P(Y)$, será um valor entre 0 (não existe possibilidade de que a saída ocorra) e 1 (a saída ocorrerá). Vimos na regressão múltipla (consulte o capitulo anterior) que para avaliar se um modelo adere aos dados, comparamos os valores observados e os previstos (se você lembrar, utilizamos R^2, que é a correlação de Pearson ao quadrado entre os valores observados de saída e os valores previstos pelo modelo de regressão). Da mesma forma, na regressão logística, podemos utilizar os valores observados e previstos para avaliar a aderência do modelo. Para fazer isso, utilizamos a **verossimilhança-log (VL)**:

$$\text{verossimilhança-log} = \sum_{i=1}^{N} \{Y_i \ln(P(Y_i)) + (1 - Y_i) \ln[1 - P(Y_i)]\} \quad (6.5)$$

Material com direitos autorais

A verossimilhança-log é, portanto, baseada na soma das probabilidades associadas com a saída real e a prevista (ver Tabachnick e Fidell, 2001). A estatística de verossimilhança-log é análoga à soma dos resíduos ao quadrado na regressão múltipla, no sentido de que ela é um indicador de quanta informação não explicada ainda existe após o modelo ter sido ajustado. Como conseqüência, tem-se que valores altos da estatística de verossimilhança-log indicam uma aderência pobre do modelo, porque quanto maior for esse valor, mais observações não explicadas existirão.

É possível calcular a verossimilhança-log para diferentes modelos e comparar esses modelos por meio das diferenças entre os resultados da estatística de verossimilhança-log. Um uso para isso é comparar o estado do modelo de regressão contra algum tipo básico de estado. O estado básico utilizado é normalmente o modelo quando apenas a constante está incluída. Na regressão múltipla, o modelo básico utilizado foi a média de todos os valores da variável Y (ela era a nossa melhor estimativa quando nenhuma outra informação estava disponível). Na regressão logística, se quiséssemos prever um resultado, qual seria o nosso melhor palpite? Bem, não podemos utilizar os valores de saída porque eles são formados por 0s e 1s e, assim, a média não tem sentido! Contudo, se soubermos a frequência de 0s e 1s, nosso melhor palpite seria a categoria com o maior número de casos. Dessa forma, se um valor de saída ocorre 107 vezes e não ocorre em 72 vezes, nossa melhor estimativa seria de que ele irá ocorrer (porque temos 107 ocorrências contra 72 não-ocorrências). Como tal, em uma regressão múltipla, nosso modelo base nos dá a melhor previsão na falta de qualquer outra informação: na regressão logística, essa mesma situação seria prever a saída que ocorre com maior freqüência. Esse é o modelo de regressão logística quando apenas a constante for incluída. Se adicionarmos um ou mais previsores ao modelo, podemos calcular a melhoria do modelo da seguinte forma:

$$\chi^2 = 2[\text{VL(Novo)} - \text{VL(Básico)}] \quad (6.6)$$
$$(gl = k_{\text{Novo}} - k_{\text{Básico}})$$

Assim, simplesmente pegamos o novo modelo e o subtraímos do modelo básico (o modelo quando somente a constante está presente). Você deve ter notado que multiplicamos esse valor por 2; fazemos isso porque ele é um resultado da distribuição qui-quadrado (veja o Capítulo 16 e o Apêndice A4) e, dessa forma, fica fácil calcular a significância do valor. A distribuição qui-quadrado que utilizamos tem graus de liberdade igual ao número de parâmetros no novo modelo menos o número de parâmetros no modelo básico. O número de parâmetros, k, no modelo básico é sempre igual a 1 (a constante é o único parâmetro a ser estimado) e qualquer modelo subseqüente terá um número de graus de liberdade igual ao número de previsores mais 1 (isto é, o número de previsores mais o parâmetro representando a constante).

6.3.2 Avaliando o modelo: R e R^2 ③

Quando falamos sobre regressão linear, vimos que o coeficiente de correlação múltiplo R e o correspondente R^2 foram medidas úteis para avaliar como o modelo se ajustava aos dados. Vimos também que a razão de verossimilhança é semelhante no sentido de que ela está baseada no nível de correspondência entre os valores de saídas previstos e observados. Contudo, é possível calcular uma versão mais adequada da correlação múltipla na regressão logística por um valor conhecido como estatística-R. Essa estatística-R é a correlação parcial entre a variável de saída e cada uma das variáveis previsoras e podem variar de −1 a 1. Um valor positivo indica que quando a variável previsora aumenta, também aumenta a probabilidade da ocorrência do evento. Um valor negativo indica que se a variável previsora aumenta, a probabilidade do valor de saída ocorrer diminui. Se uma variável tem

Material com direitos autorais

um valor pequeno de R, então ela contribui pouco para o modelo.

O cálculo de R é fornecido na equação (6.7). O -2VL é o -2 verossimilhança-log para o modelo original, a estatística de Wald é calculada como o descrito na próxima seção e os graus de liberdade podem ser lidos na tabela resumo das variáveis na equação. Contudo, em virtude do valor de R ser dependente da estatística de Wald, ele não é uma medida precisa (veremos na próxima seção que a estatística de Wald pode não ser precisa em certas circunstâncias). Por esse motivo, o valor de R deve ser tratado com alguma cautela e é inválido elevá-lo ao quadrado e interpretá-lo como foi feito na regressão linear.

$$R = \pm\sqrt{\frac{\text{Wald} - (2 \times gl)}{-2\text{VL(Original)}}} \qquad (6.7)$$

Existe controvérsia sobre o que seria um bom valor semelhante ao R^2 da regressão linear, mas uma medida descrita por Hosmer e Lemeshow (1989) pode ser facilmente calculada. Na terminologia do SPSS, o R^2_V de **Hosmer e Lemeshow** é calculado como na equação (6.8):

$$\text{R}^2_\text{V} = \frac{-2\text{VL(Modelo)}}{-2\text{VL(Original)}} \qquad (6.8)$$

Como tal, o R^2_V é calculado pela divisão do qui-quadrado do modelo (baseado na verossimilhança-log) pelo -2VL-*original* (a verossimilhança-log para o modelo antes que qualquer previsor tenha sido adicionado). O R^2_V é a redução proporcional no valor absoluto da medida verossimilhança-log e, desse modo, é uma medida de quanto a não-aderência aumenta como resultado da inclusão de uma variável previsora. Ele pode variar de 0 (indicando que os previsores são inúteis na previsão da variável de saída) a 1 (indicando que o modelo prevê perfeitamente a variável de saída).

Contudo, essa não é a medida utilizada pelo SPSS. O SPSS utiliza o R^2_{CS} de **Cox e Snell** (1989), que é baseado na verossimilhança-log do modelo (VL(Novo)) e a verossimilhança-log do modelo original (VL(Básico)) e o tamanho da amostra, n:

$$R^2_{CS} = 1 - e^{\left[-\frac{2}{n}(\text{VL(Novo)} - \text{VL(Básico)})\right]} \qquad (6.9)$$

Contudo, essa estatística nunca alcança o seu valor teórico máximo, 1. Portanto, Nagelkerke (1991) sugeriu a seguinte correção (R^2_N de **Nagelkerke**):

$$R^2_\text{N} = \frac{R^2_{CS}}{1 - e^{\left[\frac{2(\text{VL(Básico)})}{n}\right]}} \qquad (6.10)$$

Embora todas essas medidas tenham diferenças na forma de cálculo (e nos resultados), coletivamente é possível considerá-las como praticamente as mesmas. Assim, em termos de interpretação elas podem ser vistas como similares ao R^2 da regressão linear no sentido de fornecer uma medida do grau de aderência do modelo.

6.3.3 Avaliando a contribuição dos previsores: a estatística de Wald ②

Como na regressão linear, queremos saber não apenas se o modelo se ajusta bem aos dados, mas também a contribuição individual de cada um dos previsores. Na regressão linear, utilizamos os coeficientes estimados da regressão (b) e seus erros padrão para calcular a estatística t. Na regressão logística, existe uma estatística análoga a de Wald que apresenta uma distribuição especial conhecida como **qui-quadrado**. Da mesma forma que o teste t na regressão linear, a estatística de Wald nos informa se o coeficiente b de cada previsor é significativamente diferente de zero. Se isso ocorrer, poderemos assumir que o previsor está contribuindo de forma significativa para a previsão da variável de saída (Y). A equação (6.11) mostra como a estatística de Wald é calculada e é possível ver que basicamente ela é igual à estatística t na regressão linear (veja a equação (5.6)); ela é o valor do coeficiente de regressão dividido pelo seu erro padrão associado. A estatística de Wald é geralmente utilizada para determinar se uma variável é um previsor signi-

Material com direitos autorais

ficativo da saída; contudo, ela é provavelmente mais precisa para examinar a estatística da razão de verossimilhança. A estatística de Wald (Figura 6.1) deve ser vista com cautela porque quando o coeficiente de regressão (b) é grande, o erro padrão tende a ficar inflacionado, resultando em uma estatística de Wald subestimada (veja Menard, 1995). A inflação do erro padrão aumenta a probabilidade de que o previsor seja rejeitado quando na realidade ele contribui para o modelo (isto é, você tem uma probabilidade maior de cometer o erro do Tipo II).

$$\text{Wald} = \frac{b}{\text{EP}_b} \tag{6.11}$$

6.3.4 Exp *b* ③

O mais importante para a *interpretação* da regressão logística é o valor da exp b (*Exp(B)* na saída do SPSS), é um indicador da mudança nas probabilidades resultantes da mudança de uma unidade no previsor. Ele é similar aos coeficientes b da regressão logística, mas mais fácil de entender (porque não requer uma transformação logarítmica). Quando a variável previsora é categórica, o exp b é mais fácil de explicar; assim, imagine que temos um exemplo simples no qual estamos tentando prever a possibilidade de gravidez de uma mulher a partir da informação do uso ou não de preservativo pelo parceiro com quem ela se relacionou. A chance de um evento ocorrer é definida como a probabilidade dele ocorrer dividida pela probabilidade dele não ocorrer (veja a equação 6.12) e não deve ser confundida com a forma mais coloquial da palavra que se refere à probabilidade. Assim, por exemplo, a chance de engravidar é a probabilidade de engravidar dividida pela probabilidade de não engravidar:

$$\text{Chance} = \frac{P(\text{evento ocorrer})}{P(\text{evento não ocorrer})}$$

$$P(\text{evento } Y) = \frac{1}{1 + e^{-(b_0 + b_1 X)}} \tag{6.12}$$

$$P\begin{pmatrix}\text{evento } Y \\ \text{não ocorrer}\end{pmatrix} = 1 - P\begin{pmatrix}\text{evento } Y \\ \text{ocorrer}\end{pmatrix}$$

Para calcular a mudança na chance que resulta da mudança de uma unidade no previsor, primeiro precisamos calcular as chances de engravidar quando não foi utilizado preservativo. Depois calculamos as chances de engravidar quando foi usado preservativo. Finalmente, calculamos a proporção entre essas duas chances.

Para calcular o primeiro conjunto de chances, precisamos utilizar a equação (6.3) a fim de calcular a probabilidade de engravidar quando não foi utilizado preservativo. Se tivéssemos mais do que um previsor, usaríamos a equação (6.4). Existem três quantidades desconhecidas nessa equação: o coeficiente da constante (b_0), o coeficiente do previsor (b_1) e o valor do próprio previsor (X). Saberemos o valor de X a partir de como foi codificada a variável de utilização do preservativo (as chances são: não utilizado = 0 e utilizado = 1). Os valores de b_0 e b_1 serão estimados por nós. Podemos calcular as chances como na equação (6.12).

Em seguida, fazemos o mesmo cálculo após a variável previsora ter mudado em uma unidade. Nesse caso, em virtude do previsor ser dicotômico, precisamos calcular a chance de gravidez dado que o preservativo foi utilizado. Assim, o valor de X agora é 1 (em vez de 0).

Nós conhecemos as chances antes e depois da mudança de uma unidade na variável previsora. É apenas uma questão de calcular a

Figura 6.1 Abraham Wald escrevendo "eu não devo construir estatísticas teste propensas a ter erros padrão inflados" no quadro-negro cem vezes.

apresentar o menor impacto no ajuste do modelo aos dados será removido primeiro.

6.3.5.3 Como selecionamos um método? ②

Da mesma forma que a regressão linear (capítulo anterior), o método de regressão escolhido dependerá de vários fatores. A principal consideração é se você está testando uma teoria ou está fazendo um trabalho exploratório. Como já foi dito, algumas pessoas acreditam que os métodos passo a passo (*stepwise*) não são bons para testar teorias. Contudo, os métodos passo a passo são apropriados quando não existem pesquisas prévias que podem ser tomadas por base para testar hipóteses e em situações onde a causalidade não é de interesse e você quer somente achar um modelo para ajustar os dados (Menard, 1995; Agresti e Finlay, 1986). Também, como foi colocado para a regressão linear, se você decidir utilizar um método passo a passo, o método para trás (*backward*) é preferível ao método para frente (*forward*). Isso é em virtude do **efeito supressor**, que ocorre quando um previsor tem um efeito significativo somente quando outra variável é mantida constante. A seleção para frente é mais provável que a eliminação para trás para excluir previsores envolvidos em efeitos supressores. Desse modo, o método oferece um alto risco de se cometer o Erro do Tipo II. Em termos da estatística teste utilizada em métodos passo a passo, a estatística de Wald tende a ser pouco precisa em certas circunstâncias (veja a Seção 6.3.3), portanto, a razão de verossimilhança é melhor.

6.4 EXECUTANDO A ANÁLISE: UM EXEMPLO DE PESQUISA ②

Como meu primeiro exemplo de pesquisa, vou apresentar um de psicologia do desenvolvimento. Um bom exemplo de uma variável dicotômica é passar ou não em um teste, que em termos de psicologia do desenvolvimento é referido como tendo ou não uma determinada habilidade cognitiva. Esse exemplo está relacionado com o entendimento de regras de desenvolvimento por uma criança com base na sua idade e se a criança apresenta uma teoria da mente. Simplificando, uma regra de desenvolvimento é uma convenção de demonstração de uma emoção apropriada em uma dada situação. Por exemplo, se você ganha um presente de Natal que não gosta, a regra apropriada de comportamento é sorrir polidamente e dizer "obrigado tia Kate, eu sempre quis um repolho podre". O comportamento emocional inapropriado é começar a chorar e gritar "por que você me comprou um repolho podre, sua velha chata?". Existem algumas evidências de que crianças pequenas não têm um entendimento do comportamento apropriado e isso está conectado à teoria do desenvolvimento do cérebro (que é simplesmente a habilidade de entender o que a outra pessoa pode estar pensando).

Para esse exemplo, nossos pesquisadores estão interessados em saber se o entendimento das regras de desenvolvimento emocionais está relacionado a ter uma teoria da mente. Eles acreditam que uma criança precisa entender como outra pessoa pensa para perceber como o seu comportamento emocional irá afetar essa pessoa: se você não conseguir se imaginar no lugar da tia Katie, então você não irá entender que ela poderá ficar muito chateada se você chamá-la de velha chata. Para testar essa teoria, foi atribuída a várias crianças uma tarefa padrão de falsas crenças (uma tarefa usada para medir se alguém tem uma teoria da mente) em que elas poderiam passar ou não e suas idades em meses foram medidas. Além disso, a cada criança foi dada uma tarefa com uma regra de comportamento que elas poderiam também passar ou não. Assim, as seguintes variáveis foram mensuradas:

Material com direitos autorais

- **Saída (variável dependente):** ter um entendimento das regras de comportamento (a criança passou no teste: Sim/Não?).
- **Previsor (variável independente)**: ter ou não uma teoria da mente (a criança passou na tarefa de falsas crenças: Sim/Não?).
- **Previsor (variável independente):** Idade em meses.

Nesse experimento, existe uma variável de saída dicotômica, um previsor categórico e um previsor contínuo. O cenário é ideal para uma regressão logística.

6.4.1 A análise principal ②

Para executar uma regressão logística, a entrada dos dados deve ser feita como na regressão linear: eles são arranjados no editor de dados em três colunas (uma para cada variável). Os dados pode ser encontrados no arquivo **display.sav** na pasta do Capítulo 6 no *site* www.artmed.com.br. Ao observar o editor de dados, você deverá notar que as duas variáveis categóricas entraram como variáveis codificadas (veja a Seção 2.4.4); isto é, os números foram especificados como representantes de categorias. Para facilitar a interpretação, a variável de saída deve ser codificada como 1 (evento ocorreu) e 0 (evento não ocorreu); neste caso, 1 representa ter um entendimento das regras de comportamento e 0 representa a ausência de compreensão das regras. Para a tarefa das falsas crenças, um código semelhante foi utilizado (1 = passou na tarefa das falsas crenças e 2 = falhou na tarefa das falsas crenças). A regressão logística está localizada em *Regression* (Regressão), que é acessado pelo menu *Analyze* (Analisar): **Analyze⇒Regression⇒Binary logistic** (Analisar⇒Regressão⇒Logística Binária). Seguindo esse caminho a caixa de diálogo principal para a regressão logística será ativada conforme mostrado na Figura 6.2.

A caixa de diálogo principal é semelhante à da regressão padrão. Existe um espaço para colocar a variável dependente (variável de saída). Nesse exemplo, a saída é o entendimento ou não das regras de comportamento, assim, você simplesmente pode clicar em **display** e transferi-la para o quadro *Dependent* (Dependente) clicando em ▶. Existe também um quadro para especificar as covariáveis (as variáveis previsoras). É possível especificar os efeitos principais e as interações na regressão logística. Para especificar um efeito principal, selecione um previsor (por exemplo, **age** [idade]) e transfira essa variável para o quadro denominado *Covariates* (Covariáveis) clicando em ▶. Para entrar com uma interação, clique em mais de uma variável no lado esquerdo da caixa de diálogo (por exemplo, marque duas ou mais variáveis) e depois clique em a*b a fim de movê-las para o quadro das *Covariates* (Covariáveis). Nesse exemplo, existem apenas

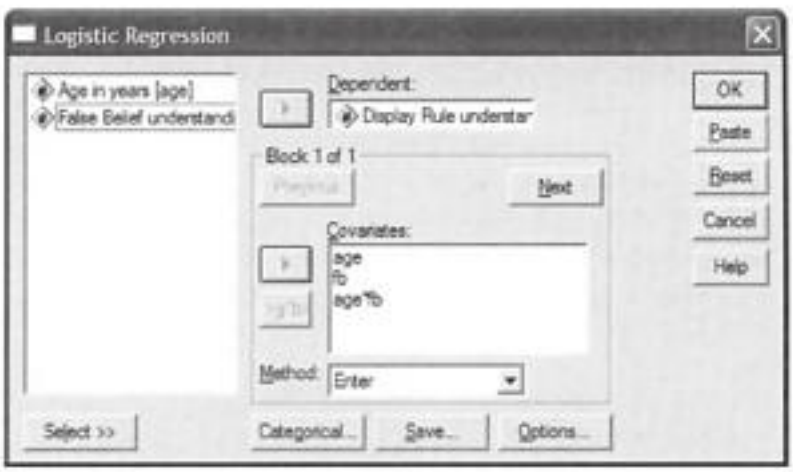


Figura 6.2 Caixa de diálogo principal da regressão logística.

Material com direitos autorais

- **Saída (variável dependente):** ter um entendimento das regras de comportamento (a criança passou no teste: Sim/Não?).
- **Previsor (variável independente)**: ter ou não uma teoria da mente (a criança passou na tarefa de falsas crenças: Sim/Não?).
- **Previsor (variável independente):** Idade em meses.

Nesse experimento, existe uma variável de saída dicotômica, um previsor categórico e um previsor contínuo. O cenário é ideal para uma regressão logística.

6.4.1 A análise principal ②

Para executar uma regressão logística, a entrada dos dados deve ser feita como na regressão linear: eles são arranjados no editor de dados em três colunas (uma para cada variável). Os dados pode ser encontrados no arquivo **display.sav** na pasta do Capítulo 6 no *site* www.artmed.com.br. Ao observar o editor de dados, você deverá notar que as duas variáveis categóricas entraram como variáveis codificadas (veja a Seção 2.4.4); isto é, os números foram especificados como representantes de categorias. Para facilitar a interpretação, a variável de saída deve ser codificada como 1 (evento ocorreu) e 0 (evento não ocorreu); neste caso, 1 representa ter um entendimento das regras de comportamento e 0 representa a ausência de compreensão das regras. Para a tarefa das falsas crenças, um código semelhante foi utilizado (1 = passou na tarefa das falsas crenças e 2 = falhou na tarefa das falsas crenças). A regressão logística está localizada em *Regression* (Regressão), que é acessado pelo menu *Analyze* (Analisar): **Analyze⇒Regression⇒Binary logistic** (Analisar⇒Regressão⇒Logística Binária). Seguindo esse caminho a caixa de diálogo principal para a regressão logística será ativada conforme mostrado na Figura 6.2.

A caixa de diálogo principal é semelhante à da regressão padrão. Existe um espaço para colocar a variável dependente (variável de saída). Nesse exemplo, a saída é o entendimento ou não das regras de comportamento, assim, você simplesmente pode clicar em **display** e transferi-la para o quadro *Dependent* (Dependente) clicando em ▶. Existe também um quadro para especificar as covariáveis (as variáveis previsoras). É possível especificar os efeitos principais e as interações na regressão logística. Para especificar um efeito principal, selecione um previsor (por exemplo, **age** [idade]) e transfira essa variável para o quadro denominado *Covariates* (Covariáveis) clicando em ▶. Para entrar com uma interação, clique em mais de uma variável no lado esquerdo da caixa de diálogo (por exemplo, marque duas ou mais variáveis) e depois clique em a*b a fim de movê-las para o quadro das *Covariates* (Covariáveis). Nesse exemplo, existem apenas

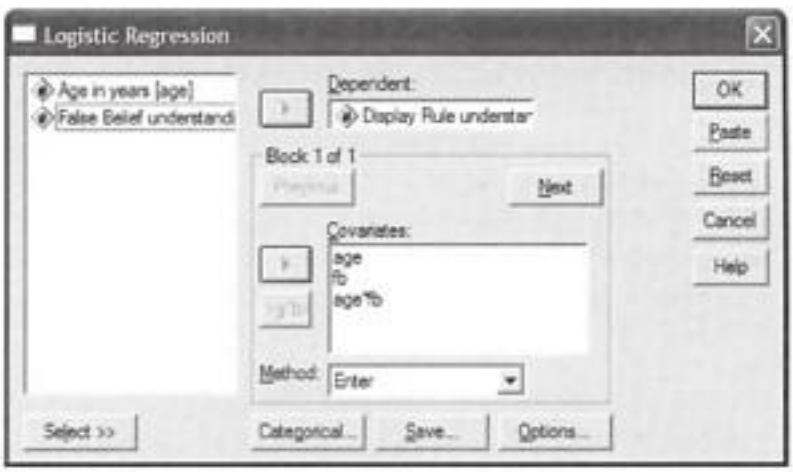


Figura 6.2 Caixa de diálogo principal da regressão logística.

Material com direitos autorais

dois previsores, portanto, existe apenas uma possível interação (a interação entre idade [**age**] × falsas crenças [**fb**]), mas se você tiver três ou mais previsores, poderá selecionar várias interações dois a dois assim como a interação das três variáveis.

6.4.2 Métodos de regressão ②

Assim como na regressão múltipla, existem vários métodos diferentes que podem ser utilizados na regressão logística (veja a Seção 6.3.5). Você pode selecionar um método específico de regressão clicando na seta que aponta para baixo próxima ao quadro denominado *Method* (Method). A Figura 6.3 mostra parte do menu da regressão logística quando os métodos de regressão estão ativos. Para essa análise, selecione o método *Forward:LR*. Depois de eu ter escrito que você nunca deveria fazer análises passo a passo, você deve estar achando um tanto estranho eu sugerir a utilização do método para a frente (forward) aqui. Bem, esse estudo é o primeiro na área e, assim, não existem pesquisas prévias que nos apontem quais previsores são confiáveis. Também, ainda não lhe mostrei um exemplo de regressão passo a passo, portanto, essa será uma maneira útil de demonstrar como o procedimento funciona!

6.4.3 Previsores categóricos ②

Nesse exemplo, existe uma variável previsora categórica. Uma das qualidades da regressão logística é que ela aceita previsores categóricos. Contudo, é necessário indicar ao SPSS quais variáveis, se existirem, são categóricas, clicando em Categorical na caixa de diálogo da regressão logística para ativar uma nova caixa de diálogo semelhante a da Figura 6.4.

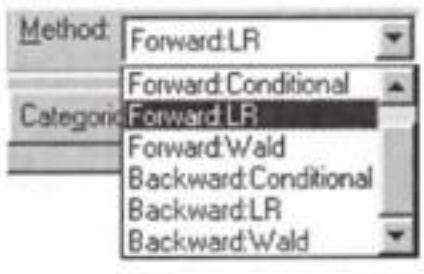


Figura 6.3 Métodos de regressão.

Nessa caixa de diálogo, as covariáveis estão listadas no lado esquerdo e existe um quadro na direita no qual as covariáveis categóricas podem ser colocadas. Destaque as variáveis categóricas que você tem (nesse exemplo, clique na **fb**) e transfira-as para o quadro *Categorical Covariates* (Covariáveis Categóricas) clicando em ▶. Na Figura 6.4, a **fb**, já está selecionada e transferida.

Existem várias maneiras de tratar previsores categóricos; por padrão, o SPSS irá utilizar os desvios dos contrastes em todas as variáveis categóricas. Na Seção 5.10 do capítulo anterior, vimos que os previsores categóricos poderiam ser incorporados na regressão pela codificação dos seus valores em zeros e uns (conhecido como código auxiliar – *dummy*). Agora existem maneiras diferentes de arranjar esses códigos dependendo do que você quer comparar. O SPSS tem várias formas padronizadas que poderão ser selecionadas (apresentaremos esses métodos em mais detalhes no Capítulo 8). Por padrão, o SPSS utiliza a codificação *Indicator* (Indicador), que é a variável auxiliar (*dummy*) padrão que já foi explicada na Seção 5.10 (e você pode escolher ainda a sua primeira ou última categoria como base). Se quiser mudar para um tipo diferente de contraste, clique na seta que aponta para baixo no quadro *Change Contrast* (Mudar contraste). A Figura 6.5 mostra que é possível selecionar contrastes simples, de diferenças, de Helmert, repetidos, polinomiais e contrastes de desvios

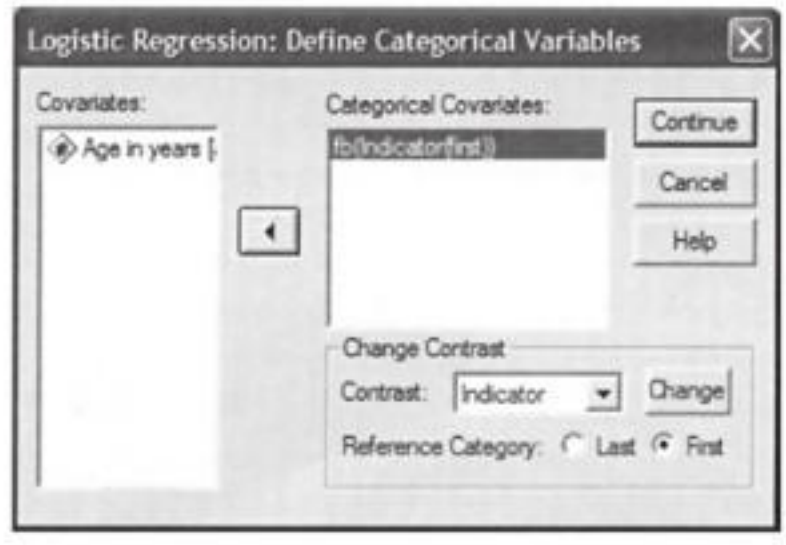


Figura 6.4 Definindo as variáveis categóricas na regressão logística.

Figura 6.5 Selecionando contrastes para previsores categóricos.

da mesma forma que na ANOVA (ver Tabela 8.6). Essas técnicas serão discutidas em detalhes no Capítulo 8, onde será explicado o que um contraste faz. Contudo, você não irá precisar dos contrastes indicadores naquele capítulo, portanto, irei utilizá-lo aqui (como na Figura 6.4). Quando um contraste indicador é utilizado, os níveis da variável categórica são recodificados utilizando uma variável *dummy* padrão (veja as Seções 5.10 e 7.8).

6.4.4 Obtendo resíduos ②

Como na regressão linear, é possível solicitar ao SPSS que determine um conjunto de resíduos como novas variáveis no editor de dados. Essas variáveis-resíduos podem então ser examinadas para verificar a qualidade do ajuste do modelo aos dados observados. Para salvar os resíduos, clique em Save na caixa de diálogo principal da regressão logística (Figura 6.2). O SPSS salva cada uma das variáveis selecionadas no editor de dados, mas elas podem ser listadas no visualizador de saídas utilizando o comando *Case Summaries* (Resumo de Casos) (veja a Seção 5.8.6) e selecionando as variáveis de interesse. A caixa de diálogo dos resíduos ilustrada na Figura 6.6 nos fornece várias opções; a maioria delas são as mesmas da regressão múltipla (veja a Seção 5.7.4). Dois tipos de resíduos exclusivos da regressão logística são as probabilidades previstas (*predicted probabilities*) e a pertinência prevista ao grupo (*predicted group memberships*). Essas probabilidades previstas são as probabilidades de Y ocorrer dado o valor de cada previsor para um determinado participante. Elas são derivadas da equação (6.4) para um dado caso. A permanência prevista ao grupo é autoexplicativa no sentido de que ela prevê a qual das duas categorias de Y o participante tem a maior probabilidade de pertencer com base no modelo. A permanência ao grupo é baseada nas probabilidades previstas e explicarei esses valores mais detalhadamente quando formos interpretar os resíduos. Vale a pena selecionar todas as opções disponíveis ou pelo menos as mesmas opções que aparecem na Figura 6.6.

O exame dos resíduos depois de qualquer análise é extremamente importante. Uma das muitas vantagens dos pacotes computacionais em relação à análise manual é que eles agilizam a obtenção de diagnósticos residuais que de outra forma seriam difíceis de conseguir devido ao tempo gasto em cálculos e pela álgebra envolvida. No entanto, o avanço computacional ainda não atingiu o ponto ótimo na consideração dos resíduos. Infelizmente, cientistas sociais, treinados a considerar apenas valores de probabilidade abaixo de 0,05 desconsideram a análise dos resíduos! Falei sobre a importância de construir modelos precisos e, também, que executar uma regressão sem conferir se o modelo se ajusta aos dados é como comprar calças sem experimentá-las – eles podem parecer boas no cabide, mas quando você as veste, em casa, ficam apertadas. As calças podem cumprir seu objetivo (elas cobrem as pernas e esquentam), mas elas não têm um valor real (porque elas cortam a circulação do sangue nas suas pernas e em outras partes importantes). Da mesma forma, a regressão executa o seu trabalho a despeito dos dados – ela irá criar um modelo – mas o valor real do modelo será limitado.

Material com direitos autorais

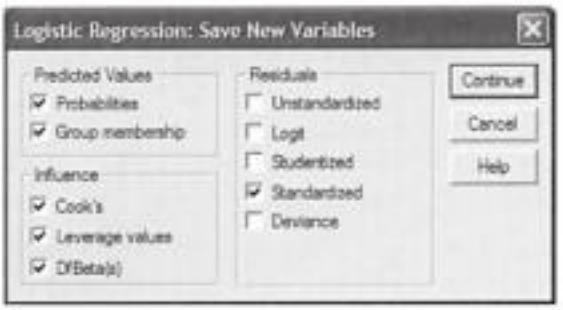


Figura 6.6 Caixa de diálogo para a determinação de resíduos na regressão.

Testar o grau de aderência com que o modelo se ajusta aos dados coletados é a essência das estatísticas de diagnóstico. Se o modelo adere bem aos dados, podemos ter uma maior segurança de que os coeficientes do modelo serão mais precisos e não serão influenciados por uns poucos pontos que se desviam. Na melhor das hipóteses, falhas no exame dos resíduos podem levá-lo a ignorar coeficientes pouco precisos no modelo. Na pior, esse modelo não acurado pode ser utilizado para aceitar hipóteses teóricas que são, de fato, falsas.

6.4.5 Opções adicionais ②

Existe uma caixa de diálogo final que oferece opções adicionais. Essa caixa é mostrada na Figura 6.7 e é acessada clicando em Options na caixa de diálogo principal da regressão logística. Em geral, as opções padrão são suficientes. Mencionei na Seção 6.4.2 que quando o método passo a passo é utilizado, existe um critério padrão para a seleção e a remoção de previsores no modelo. Esses critérios estão apresentados na caixa de diálogo *options* (opções) sob o título *Probability for Stepwise* (Probabilidade para o passo a passo). A probabilidade limite pode ser alterada, mas há necessidade, a menos que você tenha uma boa razão para querer um critério mais rígido para a seleção de variáveis. Outro valor padrão é a chegada ao modelo após um máximo de 20 iterações. Sempre que tentamos ajustar o melhor modelo possível aos nossos dados, e o número máximo de iterações pode ser imaginado como o número máximo de tentativas que o computador fará para encontrar esse melhor modelo. A menos que você tenha um modelo bastante complexo, 20 iterações serão mais do que suficientes. Vimos, no Capítulo 5, que as equações de regressão contêm uma constante que representa o intercepto com Y (isto é, o valor de Y quando os valores dos previsores são iguais a zero). Por padrão, o SPSS inclui a constante no modelo, mas é possível realizar a análise sem a constante, e isso tem o efeito de fazer o modelo passar pela origem (isto é, Y é zero quando X é zero). Dado que estamos interessados em produzir um modelo

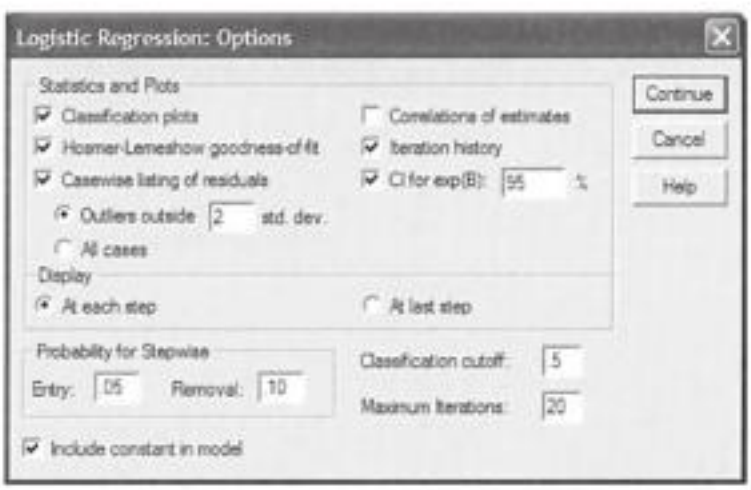


Figura 6.7 Caixa de diálogo para as opções da regressão logística.

Material com direitos autorais

Para a primeira análise, requisitamos o método passo a passo para frente (*forward stepwise*), portanto, o modelo inicial é determinado utilizando somente a constante na equação de regressão. A Saída 6.2 do SPSS nos informa sobre o modelo quando somente a constante é incluída (isto é, todas as variáveis previsoras são omitidas). Embora o SPSS não apresente esse valor, a verossimilhança-log do modelo base (veja a Seção 6.3.1) é 96,124 (acredite em mim por enquanto!). Isso representa o valor da aderência quando o modelo mais básico é ajustado aos dados. Quando inclui somente a constante, o computador cria um modelo que atribui todos os participantes a uma única categoria da variável de saída. Nesse exemplo, o SPSS pode decidir tanto prever que cada criança tem um entendimento das regras de comportamento ou prever que cada criança não tem entendimento das regras de comportamento. Ele pode tomar essa decisão arbitrariamente, mas porque é crucial maximizar quão bem o modelo prevê os valores observados, o SPSS irá prever que todas as crianças pertencem à categoria onde a maioria dos dados estaria. Nesse exemplo, 39 crianças têm um entendimento das regras de comportamento e somente 31 não. Dessa forma, se o SPSS prevê que todas as crianças têm um entendimento das regras de comportamento, essa previsão estará correta 39 vezes em 70 (isto é, 56% aproximadamente). Contudo, se o SPSS prevê que cada criança não apresenta entendimento das regras de comportamento, a previsão estará correta 31 vezes em 70 (44% aproximadamente). Assim, das duas opções, é melhor prever que todas as

Saída do SPSS 6.2

***Classificação Table*[a, b] (Tabela de classificação)**

Observed (Observado)			*Predicted* (Previstos)		
			Display Rule Understanding (Entendimento das Regras de Comportamento)		*Percentage Correct* (Percentual Correto)
			No (Não)	*Yes* (Sim)	
Step 0 (Passo 0)	*Display Rule Understanding* (Entendimento das Regras de Comportamento)	*No* (Não)	0	31	0.0
	Overall Percentage (Estatísticas Gerais)	*Yes* (Sim)	0	39	100.0 55.7

a A constante está incluída no modelo.
b O valor de corte é 0,500.

***Variables in the Equation* (Variáveis na equação)**

		B	*S. E.* (Erro Padrão)	*Wald*	*df* (gl)	*Sig.* (Sig.)	*Exp(B)*
Step 0 (Passo 0)	*Constant* (Constante)	0.230	0.241	0.910	1	0.340	1.258

***Variables not in the Equation* (Variáveis que não estão na equação)**

			Score (Escore)	*df* (gl)	*Sig.* (Sig.)
Step 0 (Passo 0)	*Variables* (Variáveis)	*AGE* (Idade)	15.956	1	0.000
		FB(1)	24.617	1	0.000
		AGE by FB(1)	23.987	1	0.000
	Overall Statistics (Estatísticas Gerais)		26.257	3	0.000

Material com direitos autorais

crianças apresentam entendimento das regras de comportamento porque isso resulta em um número maior de previsões corretas. A saída do SPSS mostra uma tabela de contingência para o modelo em seu estado básico. O SPSS previu que todas as crianças têm entendimento das regras de comportamento, o que resulta numa acurácia de 0% para as crianças observadas como não tendo entendimento das regras de comportamento, e 100% para as que passaram no teste da tarefa das regras de comportamento. No todo, o modelo classifica corretamente 55,71% das crianças. A próxima parte da saída resume o modelo, e isso requer citar o valor da constante (b_0), que é igual a 0,23.

A tabela final da saída é denominada Variáveis que não estão na Equação (*Variables not in the Equation*). A última linha dessa tabela apresenta a estatística qui-quadrado dos resíduos como 26,257, que é significativa a $p < 0,0001$ (ela é chamada de estatística global – *Overall Statistics*). Essa estatística informa que os coeficientes para as variáveis que não estão no modelo são significativamente diferentes de zero; em outras palavras, que a adição de uma ou mais dessas variáveis ao modelo irá afetar significativamente o seu poder de previsão. Se a probabilidade para o qui-quadrado residual fosse maior do que 0,05, isso implicaria que nenhuma das variáveis excluídas do modelo contribuiria de forma significativa para o poder preditivo do modelo. Desse modo, a análise terminaria nesse estágio.

O restante dessa tabela lista cada um dos previsores com o valor da **estatística do escore eficiente de Roa** para cada uma (coluna denominada Escore [*Score*]). Em amostras grandes, quando a hipótese nula é verdadeira, a estatística escore é idêntica à estatística de Wald e à estatística da razão de verossimilhança. Ela é utilizada nesse estágio da análise porque é computacionalmente menos intensiva que a estatística de Wald e, assim, pode ser calculada em situações em que esta seria proibitiva. Como qualquer estatística teste, a estatística escore de Roa tem uma distribuição específica a partir da qual a sua significância pode ser determinada. Nesse exemplo, todas as variáveis excluídas têm estatísticas escore significativas a $p < 0,001$ e, dessa forma, todas podem potencialmente contribuir para o modelo. Contudo, como mencionado na Seção 6.4.2, os cálculos passo a passo são relativos, portanto, a variável que será selecionada para inclusão é a que apresentar o maior valor para a estatística escore e que seja significante ao nível de 0,05. Nesse exemplo, tal variável é a **fb** porque ela apresenta a estatística escore de valor mais alto. A próxima parte da saída lida com o modelo após os previsores terem sido adicionados.

6.5.2 Passo 1: Entendendo falsas crenças ③

No primeiro passo, o entendimento de falsas crenças (**fb**) é adicionado ao modelo como um previsor. Assim, uma criança é classificada como tendo um entendimento das regras de comportamento com base no sucesso ou não no teste da tarefa das falsas crenças. Agora, isso pode ser facilmente explicado se olharmos a tabulação cruzada para as variáveis **fb** e **display**.[1] O modelo utilizará o entendimento de falsas crenças para prever se uma criança compreende as regras de comportamento aplicando a tabela de tabulação cruzada mostrada na Tabela 6.1. Desse modo, o modelo prevê que todas as crianças que mostraram entender falsas crenças apresentarão compreensão das regras de comportamento. Existem 41 crianças com entendimento de falsas crenças, assim, o modelo prevê que 41 crianças têm entendimento das regras de comportamento (ele estará correto 33 vezes em 41 e incorreto 8 vezes em 41). Além disso, esse novo modelo prevê que todas as 29 crianças que não apresentaram entendimento de falsas crenças não apresentam entendimento das regras de comportamento (nessa situação, ele estará correto 23 vezes em 29 e errado 6 vezes em 29).

A Saída 6.3 do SPSS mostra um resumo das estatísticas sobre o novo modelo (que,

[1] A caixa de diálogo para produzir essa tabela pode ser obtida acionando os menus **Analyze⇒Descriptive Statistics⇒Crosstabs** (Analisar⇒Estatística Descritiva⇒Tabulação cruzada).

Material com direitos autorais

Table 6.1 Tabulação cruzada das regras de comportamento com o entendimento de falsas crenças

		False Belief Understanding (Fb) (Entendimento de Falsas Crenças)	
		No (Não)	Yes (Sim)
Display Rule (Regra de Comportamento)	No (Não)	23	8
Understanding (display) (Entendimento)	Yes (Sim)	6	33

Saída do SPSS 6.3

Omnibus Tests of Model Coeficientes
(Testes em etapas dos coeficientes do modelo)

		Chi-square (Qui-quadrado)	*df* (gl)	*Sig.* (Sig.)
Step (Passo 1)	*Step* (Passo)	26.083	1	0.000
	Block (Bloco)	26.083	1	0.000
	Model (Modelo)	26.083	1	0.000

***Model Summary* (Resumo do Modelo)**

Step	*–2 Log likelihood* (–2 Verossimilhança-log)	*Cox & Snell R Square* (R quadrado de Cox e Snell)	*Nagelkerke R Square* (R quadrado de Nagelkerke.)
1	70.042	0.311	0.417

***Classification Table*[a] (Tabela de Classificação)**

			Predicted (Previstos)		
			Display Rule Understanding (Entendimento das Regras de Comportamento)		*Percentage Correct* (Percentual Correto)
Observed (Observado)			*No* (Não)	*Yes* (Sim)	
Step 1 (Passo 1)	*Display Rule Understanding* (Entendimento das Regras de Comportamento)	*No* (Não)	23	8	74.2
		Yes (Sim)	6	33	84.6
	Overall Percentage (Estatísticas Gerais)				80.0

a O valor de corte é 0,500.

como já vimos, contém **fb**). A aderência global do novo modelo é determinada utilizando a estatística de verossimilhança-log (veja a Seção 6.3.1). No SPSS, em vez de relatar o valor da verossimilhança-log, o valor é multiplicado por −2 (e algumas vezes denominado −2VL (*−2VL*)): essa multiplicação é feita porque −2VL tem uma distribuição aproximadamente qui-quadrado e, dessa forma, é possível comparar os seus valores contra aqueles que esperaríamos que fossem obtidos apenas por acaso. Lembre-se de que valores grandes da estatística VL indicam um modelo com uma má aderência.

Nesse estágio da análise, o valor de −2x verossimilhança-log deve ser menor do que quando somente a constante foi incluída no modelo (porque valores baixos de −2VL in-

Material com direitos autorais

dicam que o modelo está prevendo a variável de saída com maior precisão). Quando somente a constante foi incluída, −2VL = 96,124, mas com a adição da variável **fb**, esse valor foi reduzido para 70,042. Essa redução nos informa que o modelo agora é melhor para prever o entendimento das regras de comportamento do que antes da variável **fb** ter sido acrescentada. Você pode determinar quão bem um modelo prevê a variável de saída utilizando a estatística do modelo qui-quadrado (*model chi-square statistic*), que avalia a diferença entre o modelo atual e o modelo quando somente a constante está incluída. Vimos na Seção 6.3.1 que podemos determinar a significância da mudança em um modelo tomando a verossimilhança-log do novo modelo e subtraindo da verossimilhança-log do modelo base. A estatística do modelo qui-quadrado funciona com base nessa ideia e, dessa forma, é igual a −2VL com a variável **fb** incluída menos o valor de −2VL quando somente a constante está presente no modelo (96,124 – 70,042 = 26,083). Esse valor tem uma distribuição qui-quadrado, portanto, sua significância estatística pode ser facilmente calculada.[2] Nesse exemplo, o valor é significativo ao nível de 0,05, assim, podemos dizer que o modelo como um todo está prevendo o entendimento das regras de comportamento melhor do que quando tinha apenas a constante. A estatística do modelo qui-quadrado é análoga ao teste F para a soma dos quadrados da regressão linear (veja o Capítulo 5). O ideal seria ter um −2VL não-significativo (indicando que a quantidade de dados não explicados é mínima) e uma estatística do modelo qui-quadrado alta (indicando que o modelo incluindo os previsores é significativamente melhor do que sem os previsores). Contudo, na realidade é possível que ambas as estatísticas sejam muito significativas.

Existe uma segunda estatística denominada estatística passo (*step*), que indica a melhoria no poder preditivo do modelo a partir do último estágio. Nesse estágio existiu apenas um passo na análise, portanto, o valor da estatística de melhoria é a mesmo do qui-quadrado do modelo. Contudo, em modelos mais complexos em que existem três ou quatro estágios, esta estatística fornece uma medida da melhoria do poder preditivo do modelo desde o último passo. Seu valor é igual ao −2VL no passo atual menos o −2VL no passo anterior. Se a estatística de melhoria é significativa, ela indica que o modelo agora prevê a saída significativamente melhor do que no último passo, e numa regressão para frente (*forward*), isso pode ser entendido como uma indicação da contribuição de um previsor para o poder preditivo do modelo. De forma semelhante, a estatística bloco (*block*) fornece a mudança no −2VL desde o último bloco (para uso em análises hierárquicas ou por blocos).

Finalmente, a tabela de classificação no final da seção de saída indica quão bem o modelo prevê a inclusão em um grupo. Em virtude de o modelo estar utilizando o entendimento de falsas crenças para prever a variável de saída, essa tabela de classificação é a mesma que a Tabela 6.1. O modelo atual classifica corretamente 23 crianças que não apresentam compreensão das regras de comportamento, mas classifica errado outras 8 (isto é, ela classifica corretamente 74,19% dos casos). Para as crianças que têm compreensão das regras de comportamento, o modelo classifica corretamente 33 e erra em 6 casos (isto é, ele classifica 84,62% dos casos corretamente). A acurácia global de classificação é, dessa forma, a média ponderada desses dois valores (80%). Assim, quando somente a constante foi incluída o modelo classificou corretamente 56% das crianças, mas com a inclusão do previsor **fb**, esse valor subiu para 80%.

A próxima parte da saída (Saída do SPSS 6.4) é crucial porque ela nos informa as estimativas dos coeficientes dos previsores incluídos no modelo. Essa seção da saída nos fornece os coeficientes e estatísticas para as variáveis que foram incluídas no modelo nesse ponto (a saber, a constante e a **fb**). O valor b é o mesmo valor da regressão linear: eles são os valores que precisamos para substituir na equação (6.4) a fim

[2] Os graus de liberdade serão o número de parâmetros no novo modelo (o número de previsores mais 1; nesse caso, com um previsor, é 2) menos o número de parâmetros no modelo base (que é 1, a constante). Portanto, nesse caso, $gl = 2 - 1 = 1$.

Material com direitos autorais

Saída do SPSS 6.4

***Variables in the Equation* (Variáveis na equação)**

								95,0% C.I for EXP(B) (I.C de 95% para EXP(B))	
		B	*S. E.* (Erro Padrão)	*Wald*	*df* (gl)	*Sig.* (Sig.)	*Exp(B)*	*Lower* (Inferior)	*Upper* (Superior)
Step (Passo) 1[a]	*FB(1)*	2.761	0.605	20.856	1	0.000	15.812	4.835	51.706
	Constant (Constante)	–1.344	0.458	8.592	1	0.003	0.261		

a Variable(s) entered on Step 1: FB (Variável(is) adicionada(s) no passo 1: FB).

de identificar a probabilidade de que um caso pertença a uma determinada categoria. Vimos na regressão linear que o valor de b representa a modificação na saída resultante da mudança de uma unidade na variável previsora. A interpretação desse coeficiente na regressão logística é muito semelhante no sentido de que ele representa a chance na *logit* da variável de saída associada à alteração de uma unidade na variável previsora. A *logit* da saída é simplesmente o logaritmo natural da chance de *Y* ocorrer.

Uma estatística importante é a estatística de Wald,[3] que tem uma distribuição qui-quadrado e nos informa se o coeficiente b para o dado previsor difere de zero de forma significativa. Se isso acontece, podemos assumir que o previsor está contribuindo de modo significativo para a previsão da saída (*Y*). Vimos, na Seção 6.3.3, que a estatística de Wald deve ser utilizada com cautela porque, quando o coeficiente da regressão (*b*) é grande, o erro padrão tende a ficar inflacionado, resultando em uma subestimativa da estatística (veja Menard, 1995). Contudo, em relação a esses dados ela parece indicar que a compreensão das falsas crenças é um previsor significativo para o entendimento das regras de comportamento (observe que a significância da estatística de Wald é menor do que 0,05).

Na Seção 6.3.2, vimos que podíamos calcular um valor análogo de *R* utilizando a equação (6.7). Para esses dados, a estatística de Wald e seus *gl* podem ser vistos na Saída do SPSS 6.4 (20,856 e 1, respectivamente) e o valor original do −2VL é 96,12. Dessa forma, R pode ser calculado como:

$$R = \pm\sqrt{\left(\frac{20{,}856-(2\times1)}{96{,}124}\right)} = 0{,}4429 \qquad (6.14)$$

Na mesma seção, vimos que a medida de Hosmer e Lemeshow ($\boldsymbol{R}_{\mathrm{V}}^2$) é calculada dividindo o qui-quadrado do modelo pelo valor original de −2VL. Nesse exemplo, o qui-quadrado do modelo após todas as variáveis terem sido adicionadas é 26,083 e o valor original de −2VL (antes que qualquer variável tenha sido adicionada) foi 96,124. Assim, $\boldsymbol{R}_{\mathrm{V}}^2 = 26{,}083/96{,}124 = 0{,}271$, que é diferente do valor que seria obtido pela elevação ao quadrado do *R* determinado acima ($R^2 = 0{,}4429^2 = 0{,}196$). Na Saída 6.3 do SPSSS, foram fornecidas duas outras medidas de R^2, descritas na Seção 6.3.2. A primeira delas é a medida de Cox e Snell,[4] que o SPSS fornece como 0,311,

[3] O SPSS na realidade apresenta a estatística de Wald ao quadrado, assim, para esses dados ela seria $(2{,}761/0{,}605)^2 = 20{,}8$, como o apresentado na tabela – ver equação (6.11).

[4] Ela é calculada a partir da equação (6.9). Lembre-se de que essa equação utiliza a verossimilhança-log e que o SPSS apresenta como −2 x verossimilhança-log. VL(Nova) é, dessa forma, 70,042/-2 = −35,021 e a VL(Base) = 96,124/−2 = −48,062. O tamanho n da amostra é 70. Portanto:

$$\begin{aligned} R_{CS}^2 &= 1 - e^{\{-\frac{2}{70}[-35{,}021-(-48{,}062)]\}} \\ &= 1 - e^{-0{,}3726} \\ &= 1 - 0{,}6889 \\ &= 0{,}311 \end{aligned}$$

Material com direitos autorais

e a segunda é o valor ajustado de Negelkerke,[5] que o SPSS calcula como sendo 0,417. Como você pode ver, existe uma diferença substancial entre os dois valores!

O último item que precisamos analisar é a exp b (*Exp(B)*) na saída do SPSS, que foi descrita na Seção 6.3.4. Para calcular a mudança nas chances que resultam da variação de uma unidade no previsor, neste exemplo, precisamos primeiro calcular as chances de uma criança entender as regras de comportamento dado que ela não tem uma compreensão de segunda ordem de falsas crenças. Depois calculamos as chances de uma criança ter entendimento das regras de comportamento dado que ela compreende falsas crenças. Finalmente, calculamos a mudança proporcional nessas duas chances.

Para calcular o primeiro conjunto de chances, precisamos utilizar a equação (6.12) a fim de calcular a probabilidade de uma criança compreender regras de comportamento visto que ela falhou na tarefa das falsas crenças. O código do parâmetro no início da saída indica que a uma criança que falhou na tarefa de falsas crenças foi atribuído o código 0; dessa forma, podemos utilizar esse valor no lugar de X. Estimamos o valor de b_1 como 2,7607 (veja Variáveis na Equação na Saída 6.4 do SPSS) e a constante, que pode ser obtida na mesma tabela, é −1,3437. Podemos calcular as chances como na equação (6.15):

$$\begin{aligned} P(\text{Evento } Y) &= \frac{1}{1+e^{-(b_0+b_1X_1)}} \\ &= \frac{1}{1+e^{-[-1,3437+(2,7607\times 0)]}} \\ &= 0,2069 \\ P\begin{pmatrix}\text{Não}\\ \text{Evento } Y\end{pmatrix} &= 1-P(\text{Evento } Y) \\ &= 1-0,2069 \\ &= 0,7931 \\ \text{Chance} &= \frac{0,2069}{0,7931} \\ &= 0,2609 \end{aligned} \qquad (6.15)$$

Em seguida, fazemos o mesmo cálculo após a variável previsora ter mudado em uma unidade. Nesse caso, em virtude do previsor ser dicotômico, precisamos calcular as chances de uma criança que passou na tarefa de entendimento das regras de comportamento, dado que ela tinha passado na tarefa das falsas crenças. Assim, o valor da variável falsas crenças, X, agora é 1 (em vez de 0). Os cálculos resultantes são mostrados na equação (6.16):

$$\begin{aligned} P(\text{Evento } Y) &= \frac{1}{1+e^{-(b_0+b_1X_1)}} \\ &= \frac{1}{1+e^{-[-1,3437+(2,7607\times 1)]}} \\ &= 0,8049 \\ P\begin{pmatrix}\text{Não}\\ \text{Evento } Y\end{pmatrix} &= 1-P(\text{Evento } Y) \\ &= 1-0,8049 \\ &= 0,1951 \\ \text{Chance} &= \frac{0,8049}{0,1951} \\ &= 4,1256 \end{aligned} \qquad (6.16)$$

Sabemos as chances antes e depois da alteração de uma unidade na variável previsora. Agora, basta calcular a mudança proporcional em chances dividindo a chance após a mudança de uma unidade no previsor pela chance antes da mudança:

[5] Ela é calculada a partir da equação (6.10) – existe uma pequena diferença porque eu utilizei valores arredondados na terceira decimal:

$$\begin{aligned} R_N^2 &= \frac{0,311}{1-e^{\left[\frac{2(-48,092)}{93}\right]}} \\ &= \frac{0,311}{1-e^{-1,3732}} \\ &= \frac{0,311}{1-0,2533} \\ &= 0,416 \end{aligned}$$

Material com direitos autorais

Dica da Samanta Ferrinho

- A aderência global do modelo é mostrada pela estatística −2x verossimilhança-log e a estatística qui-quadrado associada. Se a significância do qui-quadrado é menor do que 0,05, então o modelo se ajusta de forma significativa aos dados.
- Verifique a tabela denominada "Variáveis na equação" a fim de ver quais variáveis são significativas para prever a saída.
- Para cada variável no modelo, verifique a estatística de Wald e sua significância (que deve ser menor que 0,05). Mais importante, utilize o valor do Exp(B) para a interpretação. Se o valor for maior do que 1, então quando o previsor aumenta, as chances de uma saída ocorrer aumentam. Já um valor menor do que 1 indica que quando o previsor aumenta, as chances de uma saída ocorrer diminuem. Para a interpretação acima mencionada ser confiável, o intervalo de confiança de Exp(B) não deve ultrapassar 1!
- Confira a tabela denominada "Variáveis que não estão na equação" para ver quais variáveis não prevêem de forma significativa a saída.

zero). Essa situação representa um modelo que está prevendo corretamente os dados de saída. Se, contudo, existirem muitos pontos agrupados no centro do diagrama, isso mostra que em muitos casos o modelo está prevendo uma probabilidade de ocorrência do evento de 0,50. Em outras palavras, para esses casos não existe mais do que uma chance de 50:50 de que o dado será previsto corretamente; o modelo pode prever esses casos tão bem quanto o lançamento de uma moeda (jogar cara ou coroa) o faria. Ainda, um bom modelo irá assegurar que poucos casos sejam mal classificados; nesse exemplo existem dois Ns no lado direito do modelo e um Y no lado esquerdo do modelo. Esses são casos mal classificados; quanto menos desses existirem, melhor será o modelo.

6.5.3 Listando probabilidades previstas ②

É possível listar as probabilidades esperadas da variável de saída que ocorrem com base no modelo final. Na Seção 6.4.4, vimos que o SPSS pode salvar os resíduos e também as probabilidades previstas. O SPSS salva essas probabilidades e a previsão de pertinência a um grupo como variáveis no editor de dados com os nomes PRE_1 e PGR_1, respectivamente. Essas probabilidades podem ser listadas utilizando a caixa de diálogo **Analyze⇒Reports⇒Case Summaries...** (Analisar⇒Relatar⇒Resumo de Casos...) (veja a Seção 5.8.6). A Saída do SPSS 6.7 mostra uma seleção das probabilidades previstas (porque o único previsor significativo é uma variável dicotômica, existirão somente dois valores de probabilidades previstas). Também vale a pena listar as variáveis previsoras como meio de esclarecer de onde as probabilidades previstas surgem.

Vimos, a partir do modelo, que o único previsor significativo da compreensão das regras de comportamento foi o entendimento das falsas crenças. Ele pode ter um valor tanto de 1 (passar na tarefa das falsas crenças) quanto de 0 (falhar na tarefa das falsas crenças). Se esses dois valores são colocados na equação (6.4) com os respectivos coeficientes de regressão, os dois valores das probabilidades são obtidos. De fato, calculamos esses valores como parte da equação (6.15) e da equação (6.16), e você deve ter observado que as probabilidades calculadas correspondem aos valores em PRE_1. Esses valores nos informam que quando uma criança não possui o entendimento de segunda ordem das falsas crenças (**fb** = 0, Não), existe uma probabilidade de 0,2069 de que ela passe na tarefa da compreensão das regras de comportamento – aproximadamente 21% de

Material com direitos autorais

Saída do SPSS 6.7

***Case Summaries*[a] (Resumo dos casos)**

	Case Number (Número do Caso)	*Age in years* (Idade em anos)	*False Belief Understanding* (Entendimento de Falsas Crenças)	*Display Rule Understanding* (Entendimento das Regras de Comportamento)	*Predicted probability* (Probabilidade prevista)	*Predicted group* (Grupo previsto)
1	1	24.00	Não	Não	0.20690	Não
2	5	36.00	Não	Não	0.20690	Não
3	9	34.00	Não	Sim	0.20690	Não
4	10	31.00	Não	Não	0.20690	Não
5	11	32.00	Não	Não	0.20690	Não
6	12	30.00	Sim	Sim	0.80488	Sim
7	20	26.00	Não	Não	0.20690	Não
8	21	29.00	Não	Não	0.20690	Não
9	29	45.00	Sim	Sim	0.80488	Sim
10	31	41.00	Não	Sim	0.20690	Não
11	32	32.00	Não	Não	0.20690	Não
12	43	56.00	Sim	Sim	0.80488	Sim
13	60	63.00	Não	Sim	0.20690	Não
14	66	79.00	Sim	Sim	0.80488	Sim
Total N		14	14	14	14	14

a Limitado aos primeiros 100 casos.

chance (1 em 5 crianças). Contudo, se a criança passa na tarefa das falsas crenças (**fb** = 1, Sim), existe uma probabilidade de 0,8049 de que ela irá passar na regra de comportamento – uma chance de 80,5% (4 em cada 5 crianças). Uma probabilidade de 0 indica que uma criança não tem chance de passar na tarefa do entendimento das regras de comportamento, e uma probabilidade de 1 indica que a criança irá definitivamente passar nessa regra. Portanto, os valores obtidos mostram a importância do entendimento das falsas crenças como pré-requisito para entendimento das regras de comportamento.

Assumindo que o modelo é preciso e que o entendimento das falsas crenças é significativo, podemos concluir que o entendimento das falsas crenças é o único previsor do entendimento das regras de comportamento. Além disso, a idade e a interação da idade com o entendimento das falsas crenças não prevêem de maneira alguma o entendimento das regras de comportamento. Como exercício, tente refazer essa análise utilizando o método da entrada forçada. Quais as diferenças nas conclusões?

Para ter certeza de que o modelo, de fato, é bom, é importante examinar os resíduos. Na Seção 6.4.4, vimos como o SPSS salva os vários resíduos no editor de dados. Agora podemos listar esses resíduos utilizando a caixa de diálogo **Analyze⇒Reports⇒Case Summaries...** (Analisar⇒Relatar⇒Resumo de Casos...) e interpretá-los.

6.5.4 Interpretando os resíduos ②

Os principais objetivos de examinar os resíduos na regressão logística são (1) isolar pontos em que o modelo tem pouca aderência e (2) isolar pontos que exercem uma influência indevida no modelo. Para avaliar o primeiro, vamos examinar os resíduos, especialmente os estudentizados, os padronizados e as estatísticas de desvio. Todas essas estatísticas tem uma propriedade em comum: 95% dos casos em média, em uma amostra normalmente distribuída, devem ter valores que estão entre ±1,96, e 99% dos casos devem ter valores que estão entre ± 2,58. Dessa forma, qualquer valor fora do intervalo ±3 deve ser motivo de preocupação e qualquer um fora do intervalo ±2,5 deve ser examinado com cuidado. Para avaliar a influência de casos individuais, utilizamos as estatísticas de influência, como a distância de

Material com direitos autorais

Cook (que é interpretada da mesma forma que na regressão linear: como uma medida da mudança no coeficiente de regressão se um caso é retirado do modelo). O valor do DFBeta, que é a versão padronizada da estatística de Cook, também nos informa sobre a influência de certos casos – qualquer valor maior do que 1 indica casos potencialmente influentes. Ainda, a estatística de influência de valores chapéu, que devem estar entre 0 (o caso não tem influência alguma) a 1 (o caso exerce completa influência sobre o modelo), nos informa se certos casos estão exercendo influência indevida sobre o modelo. O valor esperado da influência (*leverage*) é definido como na regressão linear. Essas estatísticas foram explicadas em mais detalhes no Capítulo 5.

Se você requisitou as estatísticas residuais, o SPSS salva cada uma como uma nova coluna no editor de dados; a Tabela 6.2 lista cada nome das variáveis e o que elas representam. Existem comentários adicionais resumindo os valores esperados de algumas estatísticas.

A Saída do SPSS 6.8 mostra as estatísticas residuais básicas para esse exemplo (distância de Cook, influência, resíduos padronizados e valores DFBeta). Note que todos os casos apresentam valores DFBetas menores do que 1 e a estatística de influência (LEV_1) próxima ao valor calculado esperado de 0,03. Também não existem valores anormalmente altos da distância de Cook (COO_1), o que significa que não existem casos influentes afetando o modelo. A distância de Cook é uma medida não-padronizada e, assim, não existe um valor absoluto que nos permita afirmar que um caso está tendo influência. Em vez disso, você deve olhar para valores da distância de Cook que são particularmente altos comparados com outros casos da amostra. Contudo, Stevens (1992) sugere que um valor maior do que 1 é problemático. Aproximadamente metade dos valores da influência são um tanto altos, mas como as outras estatísticas estão bem, isso não deve ser um motivo de preocupação.

Os resíduos padronizados têm valores menores do que $\pm 2{,}5$ e em geral apresentam valores menores do que ± 2, portanto, parece não haver motivos para se preocupar.

Você poderá notar que esses resíduos não são usuais porque são baseados em um único previsor categórico. Esse é o motivo de não existir muita variabilidade nos valores dos resíduos. Também, se valores atípicos subs-

Table 6.2 Resumo das estatísticas residuais salvas pelo SPSS

Rótulo	Nome	Comentário
PRE_1	Valor previsto	
PGR_1	Grupo previsto	
COO_1	Distância de Cook	Deve ser menor do que 1
LEV_1	Influência (*Leverage*)	Deve estar entre 0 (sem influência) e 1 (completa influência). A influência esperada é (k +1)/N, onde *k* é o número de previsores e *N* é o tamanho da amostra. Nesse caso, ela deve ser 2/70 = 0,03.
RES_1	Resíduo	
LRE_1	Resíduo Logit	
SRE_1	Resíduo Estudentizado	5% podem estar além $\pm 1{,}96$ e 1% além $\pm 2{,}58$.
ZRE_1	Resíduo padronizado	Casos acima de 3 são motivos de alerta e casos próximos de 3 merecem uma investigação.
DEV_1	Desvios	
DFB0_1	DFBeta para a constante	Deve ser menor do que 1.
DEF1_1	DFBeta para o primeiro previsor (**fb**)	

Material com direitos autorais

Saída do SPSS 6.8

Case Summaries **(Resumo dos casos)**

	Case Number (Número do Caso)	*Analog of Cook's influence statistics* (Análogo da estatística de influência de Cook.)	*Leverage value* (Valor da influência)	*Normalized residual* (Resíduo normalizado)	*DFBeta for constant* (DFBeta para a constante)	*DFBeta for FB(1)* (DFBeta para a FB(1))
1	1	0.00932	0.03448	–0.51075	–0.04503	0.04503
2	2	0.00932	0.03448	–0.51075	–0.04503	0.04503
3	3	0.00932	0.03448	–0.51075	–0.04503	0.04503
4	4	0.00932	0.03448	–0.51075	–0.04503	0.04503
5	5	0.00932	0.03448	–0.51075	–0.04503	0.04503
6	6	0.13690	0.03448	1.95789	0.17262	–0.17262
7	7	0.00932	0.03448	–0.51075	–0.04503	0.04503
8	8	0.00932	0.03448	–0.51075	–0.04503	0.04503
9	9	0.13690	0.03448	1.95789	0.17262	–0.17262
10	10	0.00932	0.03448	–0.51075	–0.04503	0.04503
11	11	0.00932	0.03448	–0.51075	–0.04503	0.04503
12	12	0.00606	0.02439	0.49237	0.00000	0.03106
13	13	0.00932	0.03448	–0.51075	–0.04503	0.04503
14	14	0.10312	0.02439	–2.03101	0.00000	–0.12812
15	15	0.00932	0.03448	–0.51075	–0.04503	0.04503
16	16	0.00932	0.03448	–0.51075	–0.04503	0.04503
17	17	0.13690	0.03448	1.95789	0.17262	–0.17262
18	18	0.00932	0.03448	–0.51075	–0.04503	0.04503
19	19	0.00606	0.02439	0.49237	0.00000	0.03106
20	20	0.00932	0.03448	–0.51075	–0.04503	0.04503
21	21	0.00932	0.03448	–0.51075	–0.04503	0.04503
22	22	0.00606	0.02439	0.49237	0.00000	0.03106
23	23	0.00606	0.02439	0.49237	0.00000	0.03106
24	24	0.10312	0.02439	–2.03101	0.00000	–0.12812
25	25	0.00932	0.03448	–0.51075	–0.04503	0.04503
26	26	0.00932	0.03448	–0.51075	–0.04503	0.04503
27	27	0.00932	0.03448	–0.51075	–0.04503	0.04503
28	28	0.00932	0.03448	–0.51075	–0.04503	0.04503
29	29	0.00606	0.02439	0.49237	0.00000	0.03106
30	30	0.00932	0.03448	–0.51075	–0.04503	0.04503
31	31	0.13690	0.03448	1.95789	0.17262	–0.17262
32	32	0.00932	0.03448	–0.51075	–0.04503	0.04503
33	33	0.00606	0.02439	0.49237	0.00000	0.03106
34	34	0.00606	0.02439	0.49237	0.00000	0.03106
35	35	0.00606	0.02439	0.49237	0.00000	0.03106
36	36	0.00606	0.02439	0.49237	0.00000	0.03106
37	37	0.10312	0.02439	–2.03101	0.00000	–0.12812
38	38	0.00606	0.02439	0.49237	0.00000	0.03106
39	39	0.13690	0.03448	1.95789	0.17262	–0.17262
40	40	0.10312	0.02439	–2.03101	0.00000	–0.12812
41	41	0.00606	0.02439	0.49237	0.00000	0.03106
42	42	0.00606	0.02439	0.49237	0.00000	0.03106
43	43	0.00606	0.02439	0.49237	0.00000	0.03106
44	44	0.00606	0.02439	0.49237	0.00000	0.03106
45	45	0.00606	0.02439	0.49237	0.00000	0.03106
Total N		45	45	45	45	45

a Limitado aos primeiros 100 casos.

Material com direitos autorais

***Case Summaries*[a] (Resumo dos casos)**

	Case Number (Número do Caso)	*Analog of Cook's influence statistics* (Análogo da estatística de influência de Cook.)	*Leverage value* (Valor da influência)	*Normalized residual* (Resíduo normalizado)	*DFBeta for constant* (DFBeta para a constante)	*DFBeta for FB(1)* (DFBeta para a FB(1))
1	46	0.10312	0.02439	–2.03101	0.00000	–0.12812
2	47	0.00606	0.02439	0.49237	0.00000	0.03106
3	48	0.00932	0.03448	–0.51075	–0.04503	0.04503
4	49	0.00932	0.03448	–0.51075	–0.04503	0.04503
5	50	0.10312	0.02439	–2.03101	0.00000	–0.12812
6	51	0.00606	0.02439	0.49237	0.00000	0.03106
7	52	0.00606	0.02439	0.49237	0.00000	0.03106
8	53	0.00606	0.02439	0.49237	0.00000	0.03106
9	54	0.00606	0.02439	0.49237	0.00000	0.03106
10	55	0.00606	0.02439	0.49237	0.00000	0.03106
11	56	0.00606	0.02439	0.49237	0.00000	0.03106
12	57	0.10312	0.02439	–2.03101	0.00000	–0.12812
13	58	0.00606	0.02439	0.49237	0.00000	0.03106
14	59	0.00606	0.02439	0.49237	0.00000	0.03106
15	60	0.13690	0.03448	1.95789	0.17262	–0.17262
16	61	0.00606	0.02439	0.49237	0.00000	0.03106
17	62	0.00606	0.02439	0.49237	0.00000	0.03106
18	63	0.00606	0.02439	0.49237	0.00000	0.03106
19	64	0.00606	0.02439	0.49237	0.00000	0.03106
20	65	0.00606	0.02439	0.49237	0.00000	0.03106
21	66	0.00606	0.02439	0.49237	0.00000	0.03106
22	67	0.00606	0.02439	0.49237	0.00000	0.03106
23	68	0.10312	0.02439	–2.03101	0.00000	–0.12812
24	69	0.00606	0.02439	0.49237	0.00000	0.03106
25	70	0.00606	0.02439	0.49237	0.00000	0.03106
Total N		25	25	25	25	25

a Limitado aos primeiros 100 casos.

tanciais ou casos influentes tivessem sido isolados, você não deveria simplesmente eliminar esses casos para melhorar o ajuste do modelo. Eles deveriam ser examinados bem de perto para tentar descobrir uma boa razão de por que eles são diferentes. Eles podem ser um erro de entrada dos dados ou uma situação especial: por exemplo, se a criança teve difi-

Dica da Samanta Ferrinho

Você precisa examinar casos que podem influenciar o modelo de regressão logística.

- Observe os resíduos padronizados: não mais do que 5% dos casos devem ter valores absolutos acima de 2 e não mais do que 1% devem ter valores absolutos acima de 2,5. Qualquer caso com um valor acima de 3 poderá ser um valor atípico.
- Procure os valores da distância de Cook no editor de dados: qualquer valor acima de 1 indica um caso que pode estar influenciando o modelo.
- Calcule a influência média (o número de previsores mais um, dividido pelo tamanho da amostra) e procure por valores maiores do que duas ou três vezes a influência média.
- Procure por valores absolutos do DFBeta maiores do que 1.

Material com direitos autorais

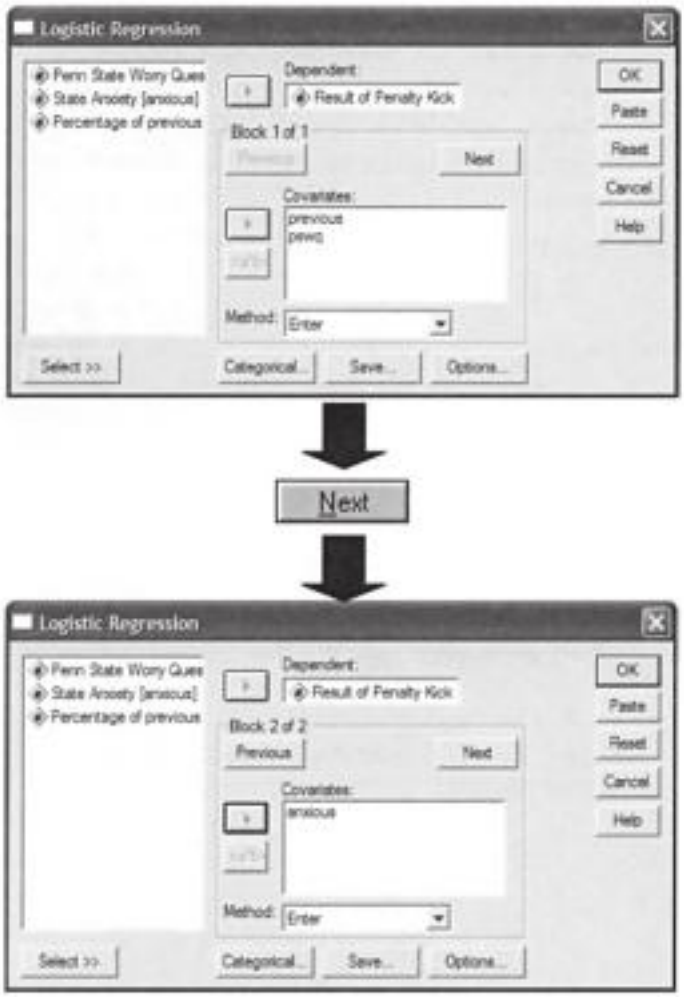


Figura 6.8 Método da entrada por blocos na regressão logística.

É fácil realizar uma regressão com entradas em blocos. Primeiro você deve utilizar o mouse para selecionar a variável **scored** (convertido) da lista de variáveis e depois transferi-la para o painel denominado *Dependent* (Dependente) clicando em . Segundo, você deve selecionar os dois previsores determinados anteriormente. Assim, selecione **pswq** e **previous** (anterior) da lista de variáveis e transfira-as para o painel denominado *Covariates* (Covariáveis) clicando em . Nosso primeiro bloco de variáveis está agora especificado. Para especificar o segundo bloco, clique em Next (próximo) a fim de limpar o painel das covariáveis, que deve estar rotulado agora como *Block 2 of 2* (Bloco 2 de 2). Selecione a variável **anxious** (ansiedade) da lista de variáveis e faça a transferência para o painel denominado *Covariates* (Covariáveis) clicando em . Podemos, nesse estágio, selecionar algumas interações para incluir no modelo, mas a menos que exista alguma boa razão teórica para acreditar que os previsores têm interação, não existe a necessidade de fazê-lo. Verifique se o *Enter* está selecionado como método de regressão (esse é o método padrão e já deve estar selecionado).

Uma vez que as variáveis tenham sido especificadas, você deve selecionar as opções descritas nas Seções 6.4.4 e 6.4.5. Como nenhum dos previsores é categórico, não há necessidade de usar a opção **Categorical...** (Categórico). Quando você tiver selecionado as opções e os resíduos que você quer, retorne à caixa de diálogo principal da regressão logística e clique em OK.

Material com direitos autorais

6.7.2 Interpretando saídas ③

A saída da regressão logística será arranjada conforme os blocos que foram especificados. Em outras palavras, o SPSS irá produzir um modelo de regressão para as variáveis especificadas no bloco 1 e depois produzirá um segundo modelo que conterá as variáveis para os dois blocos. Primeiro, a Saída 6.9 do SPSS mostra os resultados do bloco 0: a saída nos informa que 75 casos foram aceitos e que a variável dependente foi codificada como 0 e 1 (porque essa variável foi codificada como 0 e 1 no editor de dados, esses códigos correspondem exatamente aos dados no SPSS). Somos então informados sobre as variáveis que estão na equação e as que não estão na equação. Nesse ponto, somente a constante está incluída no modelo, e sendo totalmente honesto, nenhuma dessas informações é interessante!

Os resultados para o bloco 1 são mostrados na Saída 6.10 do SPSS, e nesta análise nós forçamos o SPSS a adicionar as variáveis **previous**

Saída do SPSS 6.9

***Dependent Variable Encoding* (Codificação da Variável Dependente)**

Original Value (Valor Original)	*Internal Value* (Valor Interno)
Missed Penalty (Pênalti Perdido)	0
Scored Penalty (Pênalti Convertido)	1

***Block 0: Beginning Block* (Bloco 0: Inicial)**

***Classificação Table*[a, b] (Tabela de classificação)**

			Predicted (Previstos)		
			Result of Penalty Kick (Resultado da cobrança)		*Percentage Correct* (Percentatual Correto)
	Observed (Observado)		*Missed Penalty* (Pênalti Perdido)	*Scored Penalty* (Pênalti Convertido)	
Step 0 (Passo 0)	*Result of Penalty Kick* (Resultado da cobrança)	*Missed Penalty* (Pênalti Perdido)	0	35	0.0
			0	40	100.0
	Overall Percentage (Percentagem Global)	*Scored Penalty* (Pênalti Convertido)			53.3

a A constante está incluída no modelo.
b O valor de corte é 0,500.

***Variables in the Equation* (Variáveis na equação)**

		B	*S. E.* (Erro Padrão)	*Wald*	*df* (gl)	*Sig.* (Sig.)	*Exp(B)*
Step 0 (Passo 0)	*Constant* (Constante)	0.134	0.231	0.333	1	0.564	1.143

***Variables not in the Equation* (Variáveis fora da equação)**

			Score (Escore)	*df* (gl)	*Sig.* (Sig.)
Step 0 (Passo 0)	*Variables* (Variáveis)	*PREVIOUS* (Anterior)	34.109	1	0.000
	Overall Statistics (Estatísticas Globais)	*PSWQ*	34.193	1	0.000
			41.558	2	0.000

Material com direitos autorais

e **pswq** no modelo de regressão. Dessa forma, essa parte da saída fornece informações sobre o modelo após as variáveis **previous** e **pswq** terem sido adicionadas. Observe que o −2VL é 48,66, ou seja, uma mudança para menos do 54,98 (que é o valor dado pelo qui-quadrado do modelo). Esse valor nos informa sobre o modelo como um todo enquanto que o bloco (*block*) nos informa quanto o modelo melhorou desde o último bloco. A mudança na quantia explicada pelo modelo é significativa ($p < 0,0001$), assim, utilizar a experiência prévia e a preocupação como previsor aumenta significativamente nossa habilidade de prever o sucesso na cobrança de penalidades. Um pouco abaixo, a tabela de classificação nos mostra que 84% dos casos podem ser corretamente classificados utilizando **pswp** e **previous**.

No exemplo das regras de comportamento, o SPSS não produziu o teste de aderência de Hosmer e Lemeshow. A razão é que esse teste não pode ser calculado quando existe somente um previsor e esse previsor é uma variável categórica dicotômica! Contudo, para esse exemplo o teste pode ser calculado. A parte importante desse teste é a própria estatística teste (7,93) e sua significância (0,339). Essa estatística testa a hipótese de que os dados observados são significativamente diferentes dos valores previstos pelo modelo. Assim, nós queremos um valor não-significativo para esse teste (porque isso indicaria que o modelo não difere significativamente dos dados observados). Aqui temos um valor não-significativo, o que é um indicativo de que temos um modelo que irá prever bem os valores.

A parte da Saída 6.10 do SPSS denominada *Variables in the Equation* (Variáveis na Equação) nos informa quais os parâmetros do modelo quando **previous** e **pswq** são utilizados como previsores. Os valores de significância da estatística de Wald para cada previsor indicam que tanto **pswq** quanto **previous** podem prever significativamente o sucesso da cobrança de pênalti ($p < 0,01$). Os valores de exp b para **previous** indicam que se o percentual de pênaltis convertidos aumenta em uma unidade, as chances de converter um pênalti também aumentam (porque exp b é maior do que 1). O intervalo de confiança para esse valor varia de 1,02 a 1,11, portanto, podemos ter certeza de que o valor populacional de exp b está em algum ponto entre esses dois valores. Além disso, em virtude dos dois valores serem maiores do que 1, podemos também estar confiantes de que o relacionamento entre a variável **previous** e o sucesso em converter um pênalti encontrado na amostra é verdadeiro para toda a população de jogadores de futebol. Os valores de exp b para a variável **pswq** indica que se o nível de preocupação aumenta por um ponto ao longo da escala de preocupação da Universidade Estadual da Pensilvânia, então as chances de converter um pênalti decresce (porque exp b é menor do que 1). O intervalo de confiança para esse valor varia de 0,68 a 0,93, assim, podemos estar bem confiantes de que o valor populacional de exp b está em algum ponto entre esses dois valores. Além disso, em virtude de os dois valores serem menores do que 1, podemos ter certeza de que o relacionamento entre a variável **pswq** e o sucesso em bater um pênalti encontrado nessa amostra pode ser atribuído para toda a população de jogadores de futebol. Se tivéssemos encontrado que o intervalo de confiança variasse de um valor menor do que 1 a um maior do que 1, isso limitaria a generalização de nossos achados porque o valor de exp b na população poderia indicar tanto uma relação positiva ($Exp(B) > 1$) quanto uma negativa ($Exp(B) < 1$).

Observando o diagrama de classificação (Saída do SPSS 6.11), percebemos que muitos casos estão agrupados nos extremos do diagrama e poucos casos estão no meio do diagrama. Isso reitera o que já sabemos: o modelo classifica corretamente a maioria dos casos. Podemos, nesse ponto, calcular o R^2 (veja a Seção 6.3.2) dividindo o qui-quadrado do modelo pelo valor original de −2VL. O resultado é mostrado na equação (6.18) e significa que o modelo pode explicar 53% da variância do sucesso na conversão de penalidades (assim, aproximadamente metade do que faz um pênalti ser convertido é ainda desconhecido):

Material com direitos autorais

Saída do SPSS 6.10

***Block 1: Method = Enter* (Bloco 1: Método = Enter)**

***Omnibus Tests of Model Coeficientes* (Testes em etapas dos coeficientes do modelo)**

		Chi-square (Qui-quadrado)	*df* (gl)	*Sig.* (Sig.)
Step 1 (Passo 1)	*Step* (Passo)	54.977	2	0.000
	Block (Bloco)	54.977	2	0.000
	Model (Modelo)	54.977	2	0.000

***Model Summary* (Resumo do Modelo)**

Step (Passo)	*–2 Log likelihood* (–2 Verossimilhança-log)	*Cox & Snell R Square* (R quadrado de Cox e Snell)	*Nagelkerke R Square* (R quadrado de Nagelkerke.)
1	48.662	0.520	0.694

***Hosmer and Lemeshow Test* (Teste de Hosmer e Lemeshow)**

Step (Passo)	*Chi-square* (Qui-quadrado)	*df* (gl)	*Sig.* (Sig.)
1	7.931	7	0.339

Contingency Table for Hosmer and Lemeshow Test
(Tabela de contingência para o teste de Hosmer e Lemeshow)

		Result of Penalty Kick = Missed Penalty (Resultado da cobrança = Pênalti perdido)		*Result of Penalty Kick = Scored Penalty* (Resultado da cobrança = Pênalti convertido)		
		Observed (Observado)	*Expected* (Esperado)	*Observed* (Observado)	*Expected* (Esperado)	*Total*
Step 1 (Passo 1)	1	8	7.904	0	0.096	8
	2	8	7.779	0	0.221	8
	3	8	6.705	0	1.295	8
	4	4	5.438	4	2.562	8
	5	2	3.945	6	4.055	8
	6	2	1.820	6	6.180	8
	7	2	1.004	6	6.996	8
	8	1	0.298	7	7.702	8
	9	0	0.108	11	10.892	11

a O valor de corte é 0,500.

***Classificação Table*[a] (Tabela de classificação)**

			Predicted (Previstos)		
			Result of Penalty Kick (Resultado da cobrança)		*Percentage Correct* (Percentual Correto)
	Observed (Observado)		*Missed Penalty* (Pênalti Perdido)	*Scored Penalty* (Pênalti Convertido)	
Step 1 (Passo 1)	*Result of Penalty Kick* (Resultado da cobrança)	*Missed Penalty* (Pênalti Perdido)	30	5	85.7
		Scored Penalty (Pênalti Convertido)	7	33	82.5
	Overall Percentage (Percentagem Global)				84.0

a O valor de corte é 0,500.

(*Continua*)

Material com direitos autorais

***Variables in the Equation* (Variáveis na equação)**

		B	*S. E.* (Erro Padrão)	*Wald*	*df* (gl)	*Sig.* (Sig.)	*Exp(B)*	*95,0% C.I. for EXP(B)* (I.C. de 95% para EXP(B))	
								Lower (Inferior)	*Upper* (Superior)
Step 1[a] (Passo 1)	*PREVIOUS* (Anterior)	0.065	0.022	8.609	1	0.003	1.067	1.022	1.114
	PSWQ	–0.230	0.080	8.309	1	0.004	0.794	0.679	0.929
	Constant (Constante)	1.280	1.670	0.588	1	0.443	3.598		

a Variável(is) adicionadas no passo 1: ANTERIOR, PSWQ.

Saída do SPSS 6.11

```
        Observed Groups and Predicted Probabilities (Grupos observados e
                          probabilidades previstas)

     16 +                                                                +
        I                                                                I
        I                                                                I
F       I                                                                I
R    12 +                                                                +
E       I                                                               SI
Q       I                                                               SI
U       I                                                               SI
E    8  +                                                               S+
N       IMM                                                             SI
C       IMM                                                             SI
I       IMM                                                             SI
A    4  +MM                                                             SS+
S       IMM               M              SS                 S          SSI
        IMMM              M      M  S    SM   S           S  SS    SM  SSS   SSSI
        IMMMM  MM   M     M    MMS  M    MM   S  S        S  SM   MSM SSM  SSSSI

Predicted ------------------+--------------+---------------+----------------
(Previstos)
   Prob:0                  0.25            0.5             0.75             1
Groups:     MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
(Grupos)

      Predicted Probability is of Merbership for Scored Penalty
      (Probabilidades previstas são de inclusão no grupo de pênaltis
      convertidos)

      The CUT Value is .50 (O Ponto de Corte é 0,50)

      Symbols: M - Missed Penalty, S - Scored Penalty (Símbolos: M - Pênalti
               perdido, S – Pênalti Convertido)

      Each Symbol Represents 1 Case (Cada símbolo representa 1 caso)
```

Material com direitos autorais

$$R^2 = \frac{\text{qui-quadrado do modelo}}{\text{original} - 2\text{VL}} = \frac{54{,}977}{103{,}6358} = 0{,}53 \quad (6.18)$$

A Saída 6.12 do SPSS mostra o que acontece ao modelo quando nosso novo previsor é adicionado (**anxious**). Essa parte da saída descreve o bloco 2, que é o modelo descrito no bloco 1 com o novo previsor adicionado. Assim, começamos com o modelo que tínhamos no bloco 1 e depois adicionamos o previsor **anxious** (ansioso) a ele. O efeito de adicionar o novo previsor ao modelo é reduzir o −2VL para 47,416 (uma redução de 1,246 a partir do modelo no bloco 1 como mostrado pelo qui-quadrado do modelo e a estatística de bloco). Essa melhoria não é significativa, ou seja, o acréscimo da variável ansioso (**anxious**) ao modelo não melhorou significativamente a habilidade de prever se um pênalti irá virar ou não um gol. A tabela de classificação nos informa que o modelo está agora classificando corretamente em 85,33% dos casos. Lembre que no bloco 1,84% foi classificado corretamente, dessa forma, 1,33% a mais dos casos são agora classificados (o que não é grande coisa – examinando a tabela podemos ver que apenas um caso novo foi corretamente classificado).

A seção denominada "Variáveis na Equação" contém agora três previsores e algo interessante aconteceu: **pswq** é ainda um previsor significativo do sucesso em converter penalidades, mas a variável **previous** não é mais significativa para prever o sucesso de um batedor fazer um gol. Além disso, a ansiedade parece não ter uma contribuição significativa para a previsão do sucesso em converter pênaltis.

Por que a experiência anterior não prevê mais o sucesso na marcação do gol nem a ansiedade se a habilidade do modelo de prever o sucesso na conversão de pênaltis aumentou ligeiramente?

O diagrama de classificação (Saída 6.13 do SPSS) é semelhante ao anterior e a contribuição do **pswq** para a previsão do sucesso em marcar o gol praticamente não está alterada. O

Saída do SPSS 6.12

***Block 2: Method = Enter* (Bloco 2: Método = Enter)**

***Omnibus Tests of Model Coeficientes* (Testes em etapas dos coeficientes do modelo)**

		Chi-square (Qui-quadrado)	*Df* (gl)	*Sig.* (Sig.)
Step 1 (Passo 1)	*Step* (Passo)	1.246	1	0.264
	Block (Bloco)	1.246	1	0.264
	Model (Modelo)	56.223	3	0.000

***Model Summary* (Resumo do Modelo)**

Step (Passo)	*–2 Log likelihood* (–2 Verossimilhança-log)	*Cox & Snell R Square* (R quadrado de Cox e Snell)	*Nagelkerke R Square* (R quadrado de Nagelkerke.)
1	47.416	0.527	0.704

***Hosmer and Lemeshow Test* (Teste de Hosmer e Lemeshow)**

Step (Passo)	*Chi-square* (Qui-quadrado)	*df* (gl)	*Sig.* (Sig.)
1	9.937	7	0.192

(*Continua*)

Material com direitos autorais

cupação não seja um previsor significativo do sucesso na conversão de pênaltis, o achado anterior de que a experiência desempenha um papel importante é incorreta. Isso seria uma conclusão perigosa; vou lhe mostrar o porquê.

6.8 TESTES PARA A MULTICOLINEARIDADE ③

Na Seção 5.6.2.4 vimos como a multicolinearidade pode afetar os parâmetros de um

Saída do SPSS 6.13

```
       Observed Groups and Predicted Probabilities (Grupos observados e
                        probabilidades previstas)

    16 +                                                                +
       I                                                                I
       I                                                                I
F      I                                                                I
R      I                                                                I
E   12 +                                                                +
Q      I                                                               SI
U      I                                                               SI
E      I                                                               SI
N    8 +                                                               S+
C      IM                                                              SI
I      IM                                                              SI
A      IMM                                                            SSI
S    4 +MM             M                                 S           SS+
       IMM             M     S                          SS           SSI
       IMMMM           M     S             S          S  MS   SS     SSSI
       IMMMMMM MM  M   M     S  S    MM  MMS   S    SM   SSMSMM SSSS   M  SSSSI

Predicted ----------------+----------------+----------------+----------------
(Previstos)
   Prob.    0             .25               .5               .75             1
   (Prob.)
   Groups  MMMMMMMMMMMMMMMMMMMM MMMMMMMMMMSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
   (Grupos)
       Predicted Probability is of Membership for Scored Penalty
       (Probabilidades previstas são de inclusão no grupo de pênaltis
       convertidos)
       The CUT Value is .50 (O ponto de corte é 0,50)
       Symbols: M - Missed Penalty, S - Scored Penalty (Símbolos: M - Pênalti
              perdido, S = Pênalti Convertido)
       Each Symbol Represents 1 Case (Cada símbolo representa 1 caso)
-
   CASE Observed (Caso observado)
        SCORED         Pred       PGroup     Resid       ZResid
        (CONVERTIDO) (Previsto) (Grupo P)  (Resíduo)   (Resíduo Z)
        58 S M**       .9312     S           -.9312       -3.6790

       S - Selected (Selecionado) U - Unselected cases (Casos não-selecionados)
       ** = Misclassified cases (Casos malclassificados)

* Cases with Studentized residuals greater than 2 are listed
  (Casos com resíduos estudentizados maiores do que 2 são listados)
       The CUT Value is .50 (O ponto de corte é 0,50)
```

Material com direitos autorais

modelo de regressão. A regressão logística é igualmente suscetível ao efeito da colinearidade, portanto, é essencial testar a colinearidade numa análise de regressão logística. Infelizmente, o SPSS não tem uma opção para fazer um diagnóstico de colinearidade na regressão logística (o que pode criar a ilusão de que não é necessário testá-la!). Contudo, você pode obter estatísticas tais como as de tolerância e FIV simplesmente executando uma análise de regressão linear utilizando a mesma variável de saída e os mesmos previsores. Assim, para o exemplo atual, acesse a caixa de diálogo para a regressão linear (*regression linear*) utilizando o mouse para especificar **Analyze⇒Regression⇒Linear** (Analisar⇒Regressão⇒Linear). A caixa de diálogo completa é mostrada na Figura 6.9. É desnecessário especificar um monte de opções (estamos utilizando a técnica apenas para testar a colinearidade), mas é essencial que você clique em Statistics... (Estatísticas) e depois selecione *Collinearity diagnostics* (diagnósticos de colinearidade) na caixa de diálogo. Uma vez que você selecionou Collinearity diagnostics, desmarque todas as opções padrão, clique em Continue (continue) para voltar à caixa de diálogo da regressão linear e clique em OK para executar a análise.

Os resultados da análise de regressão linear são mostrados na Saída 6.14 do SPSS. Da primeira tabela podemos ver que os valores da tolerância são 0,014 para a variável **previous** e **anxious** e 0,575 para a **pswq**. No Capítulo 5, vimos vários critérios para determinar a colinearidade. Menard (1995) sugere que um valor da tolerância menor do que 0,1 provavelmente indica um problema sério de colinearidade. Myers (1990) também sugere que um valor FIV maior do que 10 é motivo de preocupação; nesses dados os valores estão acima de 70 tanto para **anxious** quanto para **previous**. Parece, a partir desses valores, que existe um problema de colinearidade entre as variáveis previsoras. Podemos investigar esse problema mais a fundo examinando os diagnósticos de colinearidade.

A Saída do SPSS 6.14 também mostra uma tabela denominada Diagnósticos de Colinearidade (*Collinearity Diagnostics*). Nessa tabela, são fornecidos os autovalores da matriz produto cruzado, o índice de condição e as proporções das variâncias de cada previsor. Se qualquer um dos autovalores nessa tabela é muito maior do que os demais, a matriz produto cruzado é mal condicionada, o que significa que as soluções dos parâmetros de regressão podem ser bastante afetadas por pequenas mudanças nos previsores ou na saída. Simplificando, esses valores nos dão uma idéia da precisão do nosso modelo: se os autovalores são similares, o modelo derivado tem pouca probabilidade de mudar por pequenas alterações nas variáveis mensuradas. O *indi-*

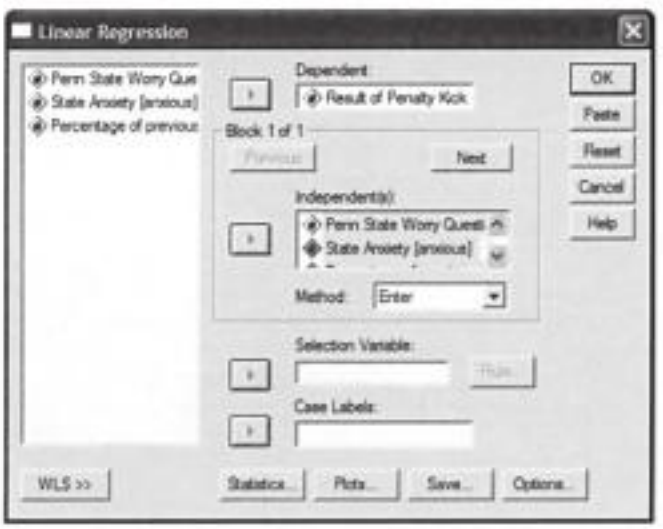


Figura 6.9 Caixa de diálogo para a *regressão linear* para dados de pênaltis.

Material com direitos autorais

Saída 6.14 do SPSS Diagnósticos de Colinearidade para os dados dos pênaltis

***Coeficients*[a] (Coeficientes)**

Model (Modelo)		Collinearity Statistics (Estatísticas de Colinearidade)	
		Tolerance (Tolerância)	*VIF*
1	*State Anxiety* (Estado da Ansiedade)	0.14	71.764
	Percentage of previous penalties scored (Percentual de pênaltis anteriores convertidos)	0.14	70.749
	Penn State Worry Questionnaire (Questionário de ansiedade da Penn State)	0.575	1.741

a. Variável dependente: Resultado do Pênalti Batido.

***Collinearity Diagnostics*[a] (Diagnósticos de Colinearidade)**

Model (Modelo)	Dimension (Dimensão)	Eingenvalue (Autovalor)	Condition Index (Índice de Condição)	Variance Proportions (Proporções da Variância)			
				(Constant) (Constante)	State Anxiety (Estado da Ansiedade)	Percentage of previous penalties scored (Percentual de pênaltis anteriores convertidos)	Penn State Worry Questionnaire (Questionário de ansiedade da Penn State)
1	1	3.434	1.000	0.00	0.00	0.00	0.01
	2	0.492	2.641	0.00	0.00	0.00	0.04
	3	7.274E-02	6.871	0.00	0.01	0.00	0.95
	4	5.195E-04	81.303	1.00	0.99	0.99	0.00

a Variável dependente: Resultado do Pênalti Batido.

ce de condição é outra forma de expressar esses autovalores, e é determinado pela raiz quadrada da razão do maior autovalor pelo autovalor de interesse (assim, para a dimensão com o maior autovalor, o índice de condição será sempre 1). Para esses dados, a dimensão final tem um índice de condição de 81,3, que é enorme se comparado com as demais dimensões. Embora não existam regras sobre o quão grande um índice de condição deva ser para indicar problemas de colinearidade, esse caso mostra claramente que o problema existe.

A etapa final da observação dessa tabela é analisar a proporção das variâncias. A variância de cada coeficiente da regressão pode ser dividida entre os autovalores, e as proporções das variâncias nos informam qual percentual da variância de cada coeficiente da regressão dos previsores é atribuído a cada autovalor. Essas proporções podem ser pensadas como percentuais para facilitar a compreensão. Assim, por exemplo, para o previsor **pswq**, 95% da variância do coeficiente da regressão está associada com o autovalor número 3, 4% está associada com o autovalor número 2 e 1% é associada ao autovalor número 1. Em termos de colinearidade, estamos procurando por previsores que apresentam altas proporções no mesmo *pequeno* autovalor, porque isso irá indicar que a variância dos seus coeficientes de regressão são dependentes. Desse modo, estamos interessados principalmente nas poucas linhas no final da tabela (que representam pequenos autovalores). Nesse exemplo, 99% da variância nos coeficientes de regressão tanto do previsor **anxiety** quanto do **previous** estão associadas ao autovalor número 4 (o menor de todos os autovalores), o qual indica, claramente, uma dependência entre essas variáveis.

O resultado dessa análise é bem claro: existe uma colinearidade entre o estado da ansiedade e a experiência prévia em bater pênaltis e essa dependência resulta em um modelo viciado. Para ilustrar de onde essa co-

Material com direitos autorais

linearidade surge, a Saída 6.15 do SPSS mostra o resultado da correlação de Pearson entre todas as variáveis nessa análise de regressão (você pode executar por si mesmo esse tipo de análise). A partir dessa saída, podemos ver que os previsores **anxious** e **previous** estão altamente correlacionados de forma negativa ($r = -0{,}99$); de fato, eles estão quase perfeitamente correlacionados. Tanto **previous** quanto **anxious** se correlacionam com o sucesso na conversão dos pênaltis[7], mas em virtude deles estarem altamente correlacionados entre si, não fica claro qual prevê o sucesso em bater um pênalti na regressão.

Essa discussão suscita a questão do que fazer quando identificamos uma colinearidade. Bem, não há muito que fazer. Uma solução lógica é retirar uma das variáveis (por exemplo, você pode ficar com o modelo do bloco 1, que ignorou o estado da ansiedade). O problema dessa opção é óbvio: não existe uma maneira de saber qual variável retirar. A conclusão teórica resultante é, dessa forma, sem sentido porque, estatisticamente falando, qualquer uma das variáveis envolvidas na colinearidade pode ser omitida. Resumindo, estatisticamente, não há problema em omitir uma variável em relação à outra. Igualmente, se um previsor é removido, Bowerman e

[7] Alguns podem questionar se é legítimo calcular a correlação de Pearson com uma variável dicotômica tal como o sucesso na conversão de um pênalti; contudo, isso é simplesmente fazer uma correlação ponto-bisserial, descrita no Capítulo 4.

Saída 6.15 do SPSS

***Correlations* (Correlações)**

		Result of Penalty Kick (Resultado de Cobrança de Pênalti)	*State Anxiety* (Estado da Ansiedade)	*Percentage of previous penalties scored* (Percentual de pênaltis anteriores convertidos)	*Penn State Worry Questionnaire* (Questionário de ansiedade da Penn State)
Result of Penalty Kick (Resultado de Cobrança de Pênalti)	*Pearson Correlation* (Correlação de Pearson)	1.000	–0.668**	0.674**	–0.675**
	Sig. (2-tailed) (Sig.) (Bilateral)	.	0.000	0.000	0.000
	N	75	75	75	75
State Anxiety (Estado da Ansiedade)	*Pearson Correlation* (Correlação de Pearson)	–0.668**	1.000	–0.993**	0.652**
	Sig. (2-tailed) (Sig.) (Bilateral)	0.000	.	0.000	0.000
	N	75	75	75	75
Percentage of previous penalties scored (Percentual de pênaltis anteriores convertidos)	*Pearson Correlation* (Correlação de Pearson)	0.674**	**–0.993****	1.000	–0.644**
	Sig. (2-tailed) (Sig.) (Bilateral)	0.000	0.000	.	0.000
	N	75	75	75	75
Penn State Worry Questionnaire (Questionário de ansiedade da Penn State	*Pearson Correlation* (Correlação de Pearson)	–0.675**	0.652**	–0.644**	1.000
	Sig. (2-tailed) (Sig.) (Bilateral)	0.000	0.000	0.000	.
	N	75	75	75	75

** A correlação é significativa ao nível de 0,01 (bilateral).

Material com direitos autorais

O'Connell (1990) afirmam que outro previsor igualmente importante que não tenha tal colinearidade forte deve substituí-lo. Eles também sugerem que se coletem mais dados para ver se a multicolinearidade pode ser reduzida. Outra possibilidade quando existem vários previsores envolvidos na multicolinearidade é executar uma análise de fatores nesses previsores e utilizar os escores dos fatores resultantes como um previsor (ver o Capítulo 15). A opção mais segura (embora insatisfatória) é reconhecer a instabilidade do modelo. Assim, se você está relatando uma análise de quais fatores prevêem o sucesso na cobrança de um pênalti, podemos reconhecer que a experiência prévia prevê de forma significativa o sucesso no primeiro modelo, mas expor que essa experiência poderá afetar a cobrança do pênalti pelo aumento do estado de ansiedade. Essa afirmação será altamente especulativa porque a correlação entre as variáveis **anxious** e **previous** nada indica sobre a direção da causalidade, mas ela valida a inexplicável conexão entre os dois previsores. Estou certo de que muitos de vocês podem julgar a falta de soluções para a colinearidade insatisfatória – infelizmente, a estatística é frustrante às vezes!

6.9 O QUE PODE DAR ERRADO ④

O SPSS resolve problemas de regressão logística por um procedimento interativo; ele supõe uma solução provável, testa para ver quão bem essa solução adere aos dados e então repete o processo para tentar melhorar o ajuste. Ele pára quando a suposição feita estiver perto o suficiente da verdadeira solução. Algumas vezes, em vez de ir rapidamente para a solução correta, você irá notar que nada acontece: o SPSS começa a se mover muito devagar. Se você não consegue encontrar uma solução correta, às vezes ele desiste oferecendo a você (sem qualquer desculpa) um resultado completamente errado. Normalmente isso é revelado por um erro padrão absurdamente grande. Duas situações podem provocar esse tipo de ocorrência.

6.9.1 Informação incompleta dos previsores ④

Imagine que você está tentando prever a ocorrência de câncer de pulmão a partir do hábito de fumar (vício que, acredita-se, aumenta o risco de câncer) e comer ou não tomates (que, acredita-se, reduz o risco de câncer). Você coleta dados de pessoas fumantes e não fumantes e de pessoas que comem e que não comem tomates; contudo, isso não é suficiente, a menos que você colete dados de todas as combinações de fumar e comer tomates. Imagine que você conseguiu os seguintes dados:

Você fuma?	Você come tomates?	Você tem câncer?
Sim	Não	Sim
Sim	Sim	Sim
Não	Não	Sim
Não	Sim	?????

Observar apenas as três primeiras possibilidades não o prepara para a saída da quarta. Não é possível saber se a última pessoa tem ou não câncer com base nos dados coletados. Dessa forma, o SPSS terá problemas a menos que você tenha coletado dados de todas as combinações das suas variáveis. Isso deve ser verificado antes de executar a análise utilizando uma tabulação cruzada; descrevo como fazer isso no Capítulo 16.

Isso se aplica não apenas para as variáveis categóricas, mas também para as contínuas. Suponha que você quer investigar fatores relacionados à felicidade. Isso poderá incluir idade, gênero, orientação sexual, crenças religiosas, níveis de ansiedade e mesmo se a pessoa é destra. Você entrevista 1000 pessoas, registra as características e se elas são ou não felizes ("sim" ou "não"). Embora a amostra de 1000 pareça bastante grande, é provável que ela inclua uma mulher budista, lésbica, canhota e altamente ansiosa de 80 anos? Se você achasse tal pessoa e ela fosse feliz, você poderia concluir que qualquer uma com as mesmas características é feliz? Seria, obviamente, necessário ter muito mais pessoas com essas

Material com direitos autorais

Imagine que nós temos a mesma almofada de pressão, se mudou e foi para a universidade. Agora estamos interessados em distinguir ladrões de gatos de estimação com base no peso. Novamente, podemos pesar alguns gatos e alguns ladrões. Dessa vez o gráfico ainda tem uma linha de triângulos em 0 (os pesos dos gatos) e uma linha 1 (os dos ladrões), mas agora as linhas dos triângulos não se sobrepõem, pois não existe um ladrão que pese o mesmo que um gato. Isso é conhecido como uma separação perfeita: a saída (gato ou ladrão) pode ser prevista perfeitamente a partir dos pesos (qualquer coisa abaixo de 15 kg é um gato e qualquer coisa acima de 40 kg é um ladrão. Se tentarmos calcular as probabilidades de uma saída dado certo peso, estaremos encrencados. Quando o peso é baixo, a probabilidade é zero, e quando é alto, a probabilidade é 1, mas o que acontece no meio? Não temos dados entre 15 e 40 em que basear essas probabilidades. A figura mostra duas curvas possíveis que podem ser ajustadas a esses dados, sendo que uma é bem mais íngreme que a outra. Qualquer uma dessas curvas é válida com base nos dados que temos. A falta de dados fará com que o SPSS não tenha certeza de quão íngreme deve ser a inclinação intermediária e ele tentará trazer o centro o mais próximo possível da vertical e estimará a mudança em direção ao infinito (com grandes erros padrão).

Esse problema às vezes ocorre quando muitas variáveis são ajustadas a poucos casos. Em muitos casos, a única solução satisfatória é coletar mais dados, mas algumas vezes uma resposta é encontrada adotando um modelo mais simples.

6.10 O QUE DESCOBRIMOS SOBRE ESTATÍSTICA? ①

Aprendi muito escrevendo este capítulo (como ignorar meu cérebro quando ele pensa "Eu sei o que seria ótimo colocar no meu livro, o capítulo de regressão logística!"). Pelo menos, espero que eu tenha conseguido disfarçar a minha incompetência pelo menos na maior parte do tempo. Começamos o capítulo entendendo por que não podemos utilizar a regressão linear (Capítulo 5) quando temos uma variável de saída dicotômica. Estudamos então um pouco de teoria sobre a regressão logística analisando a equação de regressão e seu significado. Depois, avaliamos o modelo e falamos sobre a estatística verossimilhança-log e o teste qui quadrado associado. Abordei os diferentes métodos de obter valores equivalentes ao R^2 na regressão (Hosmer e Lemeshow, Cox e Snell e Nagelkerke). Descobrimos também a estatística de Wald e o Exp(B). O restante do capítulo apresentou dois exemplos utilizando o SPSS para executar a regressão logística. Com isso, espero que você tenha compreendido de forma geral como conduzir e executar uma regressão logística.

6.11 TERMOS-CHAVE QUE DESCOBRIMOS

- –2VL
- Distribuição qui-quadrado
- Separação completa
- R^2_{CS} de Cox e Snell
- Exp(B)
- R^2_L de Hosmer e Lemeshow
- Regressão Logística
- Verossimilhança-log
- Estimação de máxima verossimilhança
- R^2_N de Nagelkerke
- Chances
- Estatística do escore eficiente de Roa
- Estatística de Wald

6.12 TAREFAS DO ALEX ESPERTO

As respostas dessas tarefas estão no *site* www.artmed.com.br no arquivo **Answers(Chapter6).pdf.**

- Pesquisas recentes mostraram que os professores estão entre os trabalhadores mais estressados. Uma pesquisadora quer saber quais os fatores que provocam estresse em um professor. Ela entrevistou 467

Material com direitos autorais

professores e utilizou vários questionários que mediram: **burnout** (colapso) (teve ou não), **perceived control** (controle percebido) (escore alto = controle percebido baixo), **coping style** (estilo de enfrentar) (escore alto = alta habilidade de enfrentar o estresse), **stress from teaching** (estresse de ensinar) (escore alto = lecionar estressa muito a pessoa), **stress from research** (estresse de pesquisar) (escore alto = pesquisar provoca muito estresse na pessoa) e **stress from providing pastoral care** (estresse de aconselhar) (escore alto = aconselhamento provoca muito estresse na pessoa). A variável de saída de interesse é **burnout** (colapso), e o modelo de estresse de Cooper, Sloan e Williams (1988) indica que o controle percebido (perceived control) e o estilo de enfrentar (coping style) são previsores importantes dessa variável. Os previsores restantes foram medidos para ver a sua contribuição sobre diferentes aspectos do trabalho do professor e seu colapso nervoso. Você pode ajudar a pesquisadora realizando uma regressão logística a fim de ver quais fatores preveem o colapso nervoso? Os dados estão no arquivo **Burnout.sav.** ③

- Um psicólogo clínico está interessado em pesquisar sobre o HIV e quer saber os fatores que influenciam o uso do preservativo com uma nova parceira (relacionamentos com menos de um mês de duração). A medida de saída é se o preservativo foi ou não-utilizado (**use** = preservativo utilizado = 1, **not used** = preservativo não-utilizado = 0). As variáveis previsoras foram principalmente resultados obtidos da utilização da EAP (Escala de Atitudes em relação ao Preservativo)* de Sacco, Levine, Reed e Thompson (*Psychological Assesment: A Journal of Consulting and Clinical Psychology, 1991*). **Gender** (Gênero – gênero da pessoa); **Safety** (Segurança – relacionamento seguro medido em uma escala de cinco pontos indicando o grau em que a pessoa vê a segurança do relacionamento em relação a doenças sexualmente transmissíveis); **Sexexp** (experiência sexual, medida em uma escala de 10 pontos indicando o grau em que experiências anteriores influenciam a atitude em relação ao uso de preservativos); **previous** (uma medida fora da escala EAP, que indica se o preservativo foi ou não utilizado pelo casal em encontros anteriores, 1 = preservativo utilizado, 0 = preservativo não utilizado, 2 = não houve encontro anterior com essa parceira); **selfcon** (autocontrole medido em uma escala de nove pontos, indicando o autocontrole da pessoa quando chega o momento de utilizar o preservativo; isto é, eles se deixaram levar pelo momento ou se controlaram?); **Perceive** (risco percebido, medido em uma escala de seis pontos, indicando o grau de preocupação da pessoa em relação ao sexo desprotegido). Pesquisas anteriores (Sacco, Rickman, Thompson, Levine e Reed, em *AIDS Education and Prevention*, 1993), mostraram que **gênero**, **relacionamento seguro** e **risco percebido** preveem o uso de preservativo. Execute uma análise apropriada para verificar esses resultados anteriores e testar se autocontrole, utilização anterior e experiência sexual podem prever qualquer variância restante na utilização de preservativo. (1) Interprete todas as partes importantes da saída do SPSS. (2) Quão confiável é o modelo final? (3) Quais são as probabilidades que os participantes 12, 53 e 75 irão utilizar preservativos? (4) Uma mulher utilizou preservativo em seu encontro anterior com um novo parceiro, e apresenta 2 em todas as variáveis exceto em risco percebido (na qual ela tem um escore de 6). Utilize o modelo para estimar a probabilidade de que ela irá utilizar preservativo em seu próximo encontro. Os dados estão no arquivo **condom.sav.** ③

* N. de T.: CAS (Condom Attitude Scale).

Material com direitos autorais

6.13 LEITURAS COMPLEMENTARES

HUTCHESON, G., SOFRONIOU, N. *The multivariate social scientist*. London: Sage, 1999. Capítulo 4.

MENARD, S. Applied logistic regression analysis. *Sage university paper series on quantitative applications in the social sciences*. Thousand Oaks (CA): Sage. 1995, pp. 07-106. Esse é um texto avançado, mas bom. Infelizmente, poucos textos básicos incluem a regressão logística, portanto, você terá que confiar no que eu escrevi!

MILES, J., SHEVLIN, M. *Applying regression and correlation: a guide for students and researchers*. London: Sage, 2001. O Capítulo 6 apresenta uma bela introdução à regressão logística.

Material com direitos autorais

7

COMPARANDO DUAS MÉDIAS

7.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ①

Em vez de olhar os relacionamentos entre variáveis, os pesquisadores às vezes estão interessados em verificar as diferenças entre grupos de pessoas. Em pesquisas experimentais em geral queremos manipular o que acontece às pessoas para que possamos fazer inferências causais. Por exemplo, se tomarmos dois grupos de pessoas e aleatoriamente designarmos um grupo a um programa de pílulas dietéticas e o outro grupo a um programa de pílulas de açúcar (que eles acham que irá auxiliá-los na perda de peso), se as pessoas que tomam as pílulas dietéticas perderem mais peso do que as que tomam pílulas de açúcar podemos inferir que as pílulas dietéticas causaram a perda de peso. Isso é uma ferramenta poderosa de pesquisa porque vai além da mera observação de variáveis e da procura por relacionamentos (como na correlação e regressão).[1] Este capítulo é o primeiro de muitos que aborda esse tipo de linha de pesquisa e começamos com o cenário mais simples: quando temos dois grupos ou, para ser mais específico, quando queremos comparar duas médias. Como veremos, existem duas maneiras de coletar dados: podemos expor diversas pessoas a diferentes manipulações experimentais (*entre grupos* ou *delineamento independente*) ou tomar um único grupo de pessoas e expô-lo a diferentes manipulações experimentais em pontos diferentes no tempo (um delineamento de *medidas repetidas*). Primeiro veremos como fazer um gráfico desses diferentes tipos de dados e, então, veremos os procedimentos estatísticos para análise de delineamentos independentes (o *teste t independente*). Depois, seguimos para o delineamento de medidas repetidas (o *teste t dependente*).

[1] As pessoas às vezes ficam confusas e pensam que certos procedimentos estatísticos permitem inferências causais e outros não. Isso não é verdade; é o fato de que em experimentos nós manipulamos uma variável sistematicamente para ver seu efeito em outra que permite a inferência causal. Em outras formas de pesquisa, meramente observamos mudanças em variáveis sem qualquer intervenção do investigador e, assim, não podemos dizer qual variável causa a mudança na outra, mas apenas que elas mudam de uma maneira específica. Como vocês irão ver os procedimentos estatísticos para analisar os diferentes tipos de dados são, na verdade, matematicamente idênticos!

Material com direitos autorais

7.2 REVISÃO DA PESQUISA EXPERIMENTAL ①

Geralmente, nas ciências sociais não estamos apenas interessados em ver quais as variáveis que covariam ou prevêem uma saída. Em vez disso, queremos olhar para o efeito de uma variável em outra pela alteração sistemática de alguns aspectos daquela variável. Assim, em vez de coletar dados que ocorrem naturalmente como na correlação e regressão, manipulamos uma variável para observar seus efeitos em outra. Como um caso simples, podemos querer ver qual o efeito que um incentivo positivo tem no aprendizado sobre estatística. Posso, portanto, dividir aleatoriamente alguns estudantes em dois grupos diferentes:

- **Grupo 1 (estímulo positivo):** Durante as aulas, parabenizo todos os estudantes nesse grupo pelo seu esforço e sucesso. Mesmo quando as pessoas fazem algo errado, eu as apóio e digo coisas como "você quase acertou, está progredindo muito bem" e dou um chocolate para o estudante.
- **Grupo 2 (estímulo negativo):** Esse grupo recebe uma aula padrão; assim, pratico abuso verbal mesmo quando os alunos fornecem a resposta correta. Não valorizo suas contribuições, sou condescendente e desdenho tudo o que dizem. Chamo os estudantes de estúpidos, ignorantes e afirmo que eles não deveriam estar fazendo esse curso.

A variável que manipulei é o método de ensino (estímulo positivo *versus* estímulo negativo). Essa variável é conhecida como *variável independente* (VI) e essa situação têm dois níveis porque ela foi manipulada de duas maneiras (isto é, o estímulo foi dividido em dois tipos: positivo e negativo). Uma vez conduzida essa manipulação, preciso ter algum tipo de saída que estou interessado em medir. Nesse caso, é a habilidade de estatística, e eu posso medir essa variável usando os resultados da prova do final do ano. Essa variável de saída é conhecida como *variável dependente* ou VD porque assumimos que esses escores irão depender do tipo do método de ensino utilizado (a **variável independente**).

7.2.1 Os dois métodos de coletar dados ①

Podemos escolher entre dois métodos de coleta de dados. O primeiro é manipular a variável independente usando participantes diferentes. Esse é o método descrito acima, em que diferentes grupos de pessoas fazem parte de cada condição experimental (*entre grupos*, *entre participantes* ou *delineamento independente*). O segundo método é manipular a variável independente usando os mesmos participantes. Simplificando, esse método significa que damos a um grupo de estudantes estímulo positivo por algumas semanas e testamos suas habilidades estatísticas e, depois, damos a esse mesmo grupo estímulo negativo por algumas semanas antes de testá-los novamente (*delineamento dentre participantes* ou *de medidas repetidas*). A maneira na qual os dados são coletados determina o tipo de teste que será utilizado para analisar os dados.

7.2.2 Dois tipos de variação ①

Imagine que você usou o delineamento de medidas repetidas num experimento que tem duas condições. Portanto, os mesmos participantes participam em ambas as condições, ou seja, medimos o comportamento dos participantes na condição 1 *e* na condição 2. Por exemplo, suponha que estamos tentando ver se conseguimos treinar chimpanzés para administrar a economia. Em uma fase do treino, eles sentam na frente de um computador e clicam teclas que mudam vários parâmetros da economia; uma vez que esses parâmetros mudaram, uma figura aparece na tela indicando o crescimento econômico resultante desses parâmetros. Como chimpanzés não sabem ler, essa realimentação não tem sentido. A segunda fase de treino é a mesma, exceto que, se

Material com direitos autorais

o crescimento da economia for bom, eles ganham uma banana (se o crescimento é ruim, eles não ganham) – essa realimentação é valiosa para o chimpanzé.

Vamos parar um pouco e pensar no que teria acontecido se **não** tivéssemos introduzido uma manipulação experimental (isto é, não houvesse bananas na segunda fase de treinamento, assim, as condições 1 e 2 seriam idênticas). Se não há manipulação experimental, esperamos que o comportamento de um chimpanzé seja o mesmo em ambas as condições. Esperamos isso porque fatores externos como idade, gênero, QI, motivação e excitação serão os mesmos em ambas as condições (o gênero do chimpanzé etc. não mudará da condição 1 para a condição 2). Se o desempenho da medição é confiável (isto é, testa quão bem eles administraram a economia) e a variável ou a característica que estamos medindo (nesse caso, a habilidade de administrar a economia) permanece estável, então o desempenho do participante na condição 1 deverá estar altamente relacionado ao seu desempenho na condição 2. Assim, chimpanzés que tiveram um alto desempenho na condição 1 terão também um alto desempenho na condição 2, e aqueles que tiveram baixo desempenho na condição 1 terão baixo desempenho na condição 2. Entretanto, o desempenho não será *idêntico*; haverá pequenas diferenças no desempenho criadas por fatores desconhecidos. Essa variação no desempenho é conhecida como *variação não-sistemática.*

Se introduzirmos uma manipulação experimental (isto é, fornecermos bananas como *feedback* em uma das seções do treinamento), diferenciamos o tratamento dos participantes da condição 1 e da condição 2. Assim, a *única* diferença entre as condições 1 e 2 é a manipulação que o investigador fez (nesse caso, os chimpanzés receberam bananas como uma recompensa positiva em uma condição, mas não na outra). Portanto, qualquer diferença entre as médias das duas condições é, provavelmente, devido à manipulação do experimento. Assim, se os chimpanzés desempenharam melhor numa fase do que na outra, isso *deve* ter ocorrido porque bananas foram usadas para fornecer um reforço ou realimentação em uma fase de treinamento e não na outra. As diferenças no desempenho criadas por uma manipulação experimental específica são conhecidas como *variação sistemática.*

Agora, vamos pensar no que acontece quando usamos participantes diferentes – um delineamento independente. Nesse delineamento ainda teremos duas condições, mas, dessa vez, participantes diferentes participam em cada condição. Voltando para o nosso exemplo, um grupo de participantes recebe treinamento sem realimentação e o segundo grupo de diferentes chimpanzés recebe reforço para o seu desempenho via bananas[2]. Imagine, novamente, que não temos uma manipulação experimental. Se nada fizéssemos com os grupos, ainda encontraríamos alguma variação de comportamento entre eles, porque diferentes chimpanzés irão variar na sua habilidade, motivação, QI e outros fatores. Resumindo, o tipo de fatores que eram mantidos constantes no delineamento de medidas repetidas está livre para variar no delineamento independente. Assim, a variação não-sistemática será maior do que para o delineamento de medidas repetidas. Se novamente introduzimos uma manipulação (isto é, bananas), mais uma vez teremos variação adicional criada por essa manipulação. Como tal, tanto no delineamento de medidas repetidas quanto no delineamento independente há sempre duas fontes de variação:

- **Variação sistemática**: Essa variação ocorre porque o investigador faz algo a todos os participantes em uma condição, mas não na outra.
- **Variação não-sistemática:** Essa variação resulta de fatores aleatórios que existem entre as condições experimentais (tais como diferenças naturais em habilidades).

[2] Quando eu digo "via", não significa que as bananas têm pequenas bocas e falaram "muito bem camarada, a economia cresceu desta vez" na linguagem de chimpanzés. O que eu quero dizer é que quando eles acertam eles recebem uma banana como recompensa pela resposta correta.

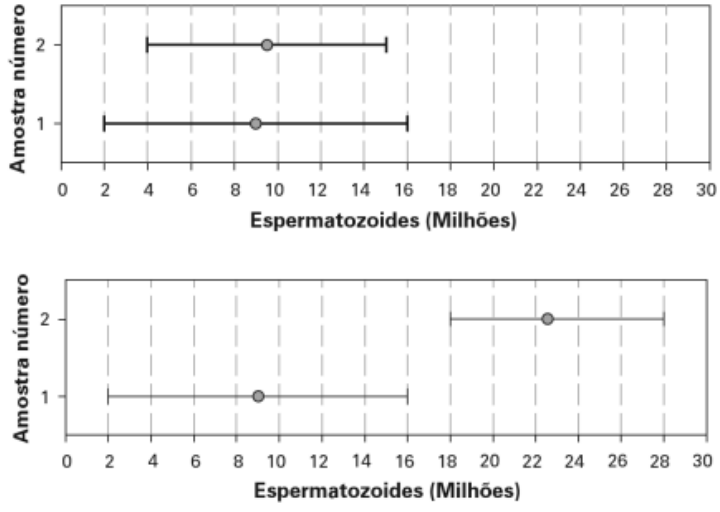
Material com direitos autorais

marmos 100 amostras de uma população, calcularmos a média de cada uma e construirmos o intervalo de confiança de 95% para cada, isso significa que 95 dos 100 intervalos de confiança conteriam o valor real da média (o valor da população). Assim, se um intervalo de confiança tem um limite inferior de 2 e um limite superior de 16, isso significa que estamos 95% confiantes que esse intervalo contém a média da população (isto é, a quantidade real de espermatozoides produzido pela codorna). Portanto, um diagrama de barra de erros mostra os limites entre os quais pensamos que o valor real da média se encontra.

Em breve você irá descobrir que o teste *t* funciona com base no princípio de que se duas amostras são tiradas da mesma população, então elas devem ter médias bem similares. Sabemos do Capítulo 1 que é possível que quaisquer duas amostras tenham médias ligeiramente diferentes (e o erro padrão nos informará até que ponto as médias podem ser diferentes). O intervalo de confiança nos diz os limites entre os quais a média da população provavelmente estará incluída (na verdade, o tamanho do intervalo de confiança dependerá do tamanho do erro padrão). No exemplo acima, sabemos que a média da população está provavelmente entre 2 e 16 milhões de espermatozoides. E se pegarmos uma segunda amostra de codornas e encontrarmos um intervalo de confiança variando de 4 a 15? Esse intervalo se sobrepõe em boa parte ao da primeira amostra:

O fato de que os intervalos de confiança se sobrepõem dessa maneira nos diz que essas médias poderiam, plausivelmente, vir da mesma população: nos dois casos estamos 95% confiantes de que os intervalos contêm o valor real da média e ambos os intervalos se sobrepõem consideravelmente, assim, eles contêm muitos valores similares. E se o intervalo de confiança para a nossa segunda amostra variasse de 18 a 28? Se compararmos isso com nossa primeira amostra, temos:

Agora esses intervalos de confiança não se sobrepõem. Assim, um intervalo de confiança (o qual estamos 95% certos que contém o valor real da média) nos diz que a média da população é algo entre 2 e 16 milhões, e o outro intervalo (o qual também estamos 95% certos que contém o valor real da média) nos informa que a média da população está entre 18 e 28. Isso sugere que ou nossas amostras vêm de diferentes populações ou ambas as amostras vêm da mesma população, mas um dos intervalos não deve conter a média da população. Se utilizarmos intervalos de 95% de confiança, saberemos que a segunda possibilidade é improvável (isso acontece somente 5



Material com direitos autorais

vezes em 100 ou 5% das vezes), assim, a primeira explicação é a mais provável.

OK, vocês devem estar pensando: "e se as amostras vêm de populações diferentes?". Bem, se pegamos duas amostras aleatórias de pessoas e as testamos em alguma medida (por exemplo, medo de livros de estatística), esperamos que essas pessoas pertençam à mesma população. Se as médias das amostras são tão diferentes para sugerir que, de fato, elas vêm de diferentes populações, como isso pode acontecer? A resposta é que nossa manipulação experimental induziu uma diferença entre as amostras.

Repetindo, quando uma manipulação experimental é bem-sucedida, esperamos que nossas amostras venham de diferentes populações. Se a manipulação não foi um sucesso, esperamos que nossas amostras venham da mesma população (isto é, as médias das amostras devem ser similares). O intervalo de 95% de confiança nos informa algo sobre o provável valor da média da população. Se tirarmos amostras de duas populações, esperamos que os intervalos de confiança sejam diferentes (para termos certeza de que as amostras são de diferentes populações, não pode haver a sobreposição dos dois intervalos de confiança). Se tirarmos duas amostras da mesma população, esperamos, se nossa medida é confiável, que os intervalos de confiança sejam muito similares (isto é, eles devem se sobrepor completamente). Você pode estar se perguntando onde esse desvio para a teoria dos testes de hipóteses está nos levando; bem, tem a ver com o que eu acabei de dizer que se nossa manipulação experimental tiver sucesso, os intervalos de confiança dos grupos-experimentais não devem se sobrepor. Se as manipulações não tiverem sucesso, os intervalos de confiança irão se sobrepor. Em termos de diagrama de barra de erros, veremos que **se as barras no nosso diagrama de barra de erros não se sobrepõem, isso indica uma diferença significativa entre grupos.** Também vale a pena mencionar, nesse momento, que diagramas de barras do SPSS são adequados somente para dados normalmente distribuídos.

7.3.1 Diagramas de barras de erros para delineamentos entre grupos ①

Quando os dados são coletados usando participantes diferentes em cada grupo, precisamos entrar com os dados usando uma variável codificadora (veja a Seção 2.4.4). Assim, o editor de dados terá duas colunas de dados. A primeira coluna é uma variável codificadora ou de código (chamada de **grupo**), a qual, quando temos somente dois grupos, irá obviamente ter dois códigos (por conveniência, sugiro 0 = grupo 1 e 1 = grupo 2). A segunda irá ter valores para a variável dependente.

Ao longo deste capítulo, uso o mesmo conjunto de dados, não porque sou muito preguiçoso para pensar em diferentes conjuntos de dados, mas porque ele me permite ilustrar vários pontos. O exemplo é se a aracnofobia (medo de aranhas) é específico a aranhas reais ou se figuras de aranhas podem evocar níveis semelhantes de ansiedade. No total, foram usados 24 aracnofóbicos. Foi pedido a 12 que brincassem com uma tarântula gigante e cabeluda, com grandes presas e um olhar maligno nos seus oito olhos. A ansiedade desse grupo foi medida. Aos 12 restantes foi apenas mostrado figuras da mesma tarântula grande e cabeluda e, novamente, sua ansiedade foi medida. Os dados estão no arquivo **spiderBG.sav**, mas para quem quiser praticar a entrada de dados, eles estão também presentes na Tabela 7.1. Lembre que cada linha no editor de dados representa um dado diferente de cada participante. Portanto, você precisa de uma coluna para designar o grupo a que eles pertencem e uma segunda coluna para representar sua ansiedade. Os dados da Tabela 7.1 mostram somente os códigos do grupo e não o rótulo correspondente. Quando você entrar os dados no SPSS, lembre de informar ao computador que o código 0 representa o grupo para o qual foi mostrada a figura e o código 1 representa o grupo que viu a aranha verdadeira (veja a Seção 2.4.4).

Depois que você entrou com os dados (ou acessou o arquivo **spiderBG.sav**), acesse a caixa de diálogo do *diagrama da barra de erros* clicando em **Graphs⇒Error Bar...** (Diagramas⇒Barra de Erros...). A caixa de

Material com direitos autorais

Tabela 7.1 Dados do **spiderGB.sav**

Participante	Grupo	Ansiedade
1	0	30
2	0	35
3	0	45
4	0	40
5	0	50
6	0	35
7	0	55
8	0	25
9	0	30
10	0	45
11	0	40
12	0	50
13	1	40
14	1	35
15	1	50
16	1	55
17	1	65
18	1	55
19	1	50
20	1	35
21	1	30
22	1	50
23	1	60
24	1	39

diálogo inicial é mostrada na Figura 7.1. Há duas escolhas do gráfico do diagrama da barra de erros. A primeira é um diagrama de barra de erro simples (para traçar níveis de uma única variável independente) e a segunda é um diagrama do gráfico de erros agrupados. O gráfico agrupado pode ser usado quando uma segunda variável independente entre grupos foi medida (tal como gênero). Nós nos limitaremos a usar um diagrama de barra de erro simples. A próxima escolha é se os dados no diagrama resumem grupos de casos ou variáveis isoladas. Há uma regra simples aqui: selecione *Summaries for groups of cases* (resumo para grupos de casos) quando os dados forem coletados usando diferentes participantes (como é o caso) e selecione *Summaries of separate variables* (resumo para variáveis isoladas) quando os dados forem coletados usando os mesmos participantes. Quando você tiver selecionado as opções apropriadas, clique em (Definir).

Figura 7.1 Caixa de diálogo inicial para diagramas de barras de erro.

Quando você clicar em (Definir), uma nova caixa de diálogo aparece (Figura 7.2), que permite especificar que variável iremos apresentar. Nesse exemplo, temos apenas duas variáveis: **group** (grupo) e **anxiety** (ansiedade). Assim, usando o mouse, selecione *Anxiety* (*anxiety*) da lista de variáveis e insira-a no espaço rotulado de *Variable* (variável) clicando em . Depois, destaque *Condition* (*group*) na lista de variáveis e transfira-a para o espaço rotulado de *Category Axis* (eixo das categorias) clicando em . Com isso você pode desenhar várias coisas usando esse tipo de gráfico, e a opção padrão é apresentar um intervalo de 95% de confiança. Este é o tipo de gráfico mais útil então as opções padrão podem ser mantidas.

O diagrama da barra de erros resultante é mostrado na Figura 7.3 e, como você pode ver, ele parece como dois Is. No meio de cada uma das duas barras há um quadrado que representa a média de cada grupo. A barra vertical mostra o intervalo de confiança em volta daquela média. Assim, a partir do gráfico desses dados pode-

Material com direitos autorais

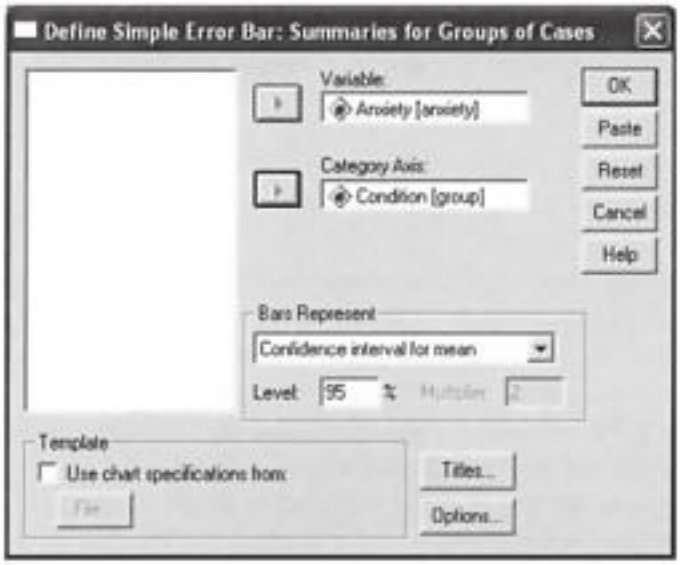


Figura 7.2 Caixa de diálogo principal para diagrama de barras de erro para resumos de grupos de casos.

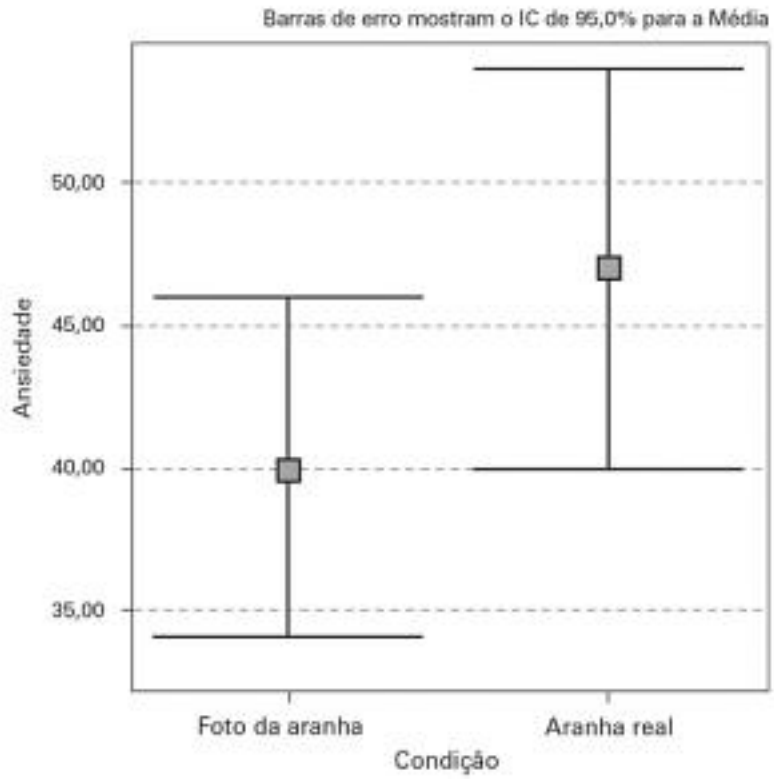


Figura 7.3 Diagrama da barra de erros de **spiderBG.sav**.

mos ver que quando a figura de uma aranha foi usada, o nível médio de ansiedade foi 40, e que provavelmente a média da população deve ficar entre 34 e 46. Entretanto, quando uma aranha verdadeira foi usada, o nível médio de ansiedade foi 47, e provavelmente a média da população deve ficar entre 40 e 54. Mais importante, as barras de erro se sobrepõem consideravelmente,

Material com direitos autorais

indicando que é improvável que essas amostras sejam de populações diferentes (portanto, a manipulação experimental não teve sucesso). Para ver se essa última afirmação é verdadeira, você terá de esperar até a Seção 7.6.3.

7.3.2 Diagramas de barras de erros para delineamentos de medidas repetidas ②

Se repetirmos o estudo recém-descrito usando os mesmos participantes em cada condição, podemos produzir um diagrama de barras de erro idêntico ao mostrado na Figura 7.3. O problema com a criação de um diagrama de barras de erro de dados de medidas repetidas é que o SPSS trata os dados como se diferentes grupos de participantes tivessem sido usados. Para provar que não estou mentindo, os dados para esse estudo estão incluídos no arquivo chamado de **spiderRM.sav.** Nesse arquivo, os valores da ansiedade são idênticos aos dados entre grupos. Entretanto, os dados estão organizados como se os mesmos participantes tivessem sido usados em cada condição (ou seja, cada participante foi exposto a uma figura de uma aranha e sua ansiedade foi medida, e em outra ocasião os mesmos participantes foram expostos a aranhas verdadeiras e sua ansiedade foi medida novamente). Os dados estão organizados de modo diferente agora porque os mesmos participantes foram usados. No SPSS, cada linha do editor de dados representa um único participante, assim, com esse desenho os dados são organizados em duas colunas (uma representando a condição da **figura da aranha** e uma representando a condição **aranha verdadeira**). Os dados estão expostos na Tabela 7.2, e eu recomendo que você tente inserir esses dados em um novo editor de dados em vez de acessar o arquivo em disco.

Para traçar o gráfico da barra de erros desses dados, selecione a caixa de diálogo da barra de erros clicando em **Graphs⇒Error Bar...** (Gráficos⇒Barra de Erros), e na caixa de diálogo (veja a Figura 7.1), clique em Summaries of separate variables (Resumos de variáveis separadas) e, depois, em Define (Definir). Esse processo mostrará a caixa de diálogo principal. Uma vez que a caixa de diálogo for ativada, você deve selecionar as duas variáveis de interesse **picture** (figura) e **real** (verdadeira) e clicar em OK. O gráfico resultante deve ser idêntico ao que você obteve para os dados entre participantes. Isso é um problema porque eu perdi muito tempo lhe dizendo como o delineamento de medidas repetidas elimina muito da variância não-sistemática dos dados e, traçando os dados das medidas repetidas dessa maneira, nós ignoramos os componentes das medidas repetidas dos dados. Loftus & Masson (1994) sugerem que uma maneira de resolver esse problema é eliminar a variabilidade entre participantes normalizando a média dos participantes (isso significa assegurar que todos os participantes tenham a mesma média entre as condições). Para normalizar as médias, precisamos usar a função *compute* (calcular) do SPSS e executar vários passos (tudo isso foi apenas um ardil para você usar a função *calcular*). Essa função já foi vista na Seção 3.3.4.

Na Seção 3.3.4, descobrimos que o comando *compute* (calcular) permite executar várias funções em colunas de dados (isto é, adicionar ou multiplicar colunas). Antes de continuarmos, retorne a essa seção e refresque sua memória sobre as maravilhas dessa função!

Tabela 7.2 Dados do **spiderRM.sav**

Sujeito	Foto (Escore da Ansiedade)	Real (Escore da Ansiedade)
1	30	40
2	35	35
3	45	50
4	40	55
5	50	65
6	35	55
7	55	50
8	25	35
9	30	30
10	45	50
11	40	60
12	50	39

Material com direitos autorais

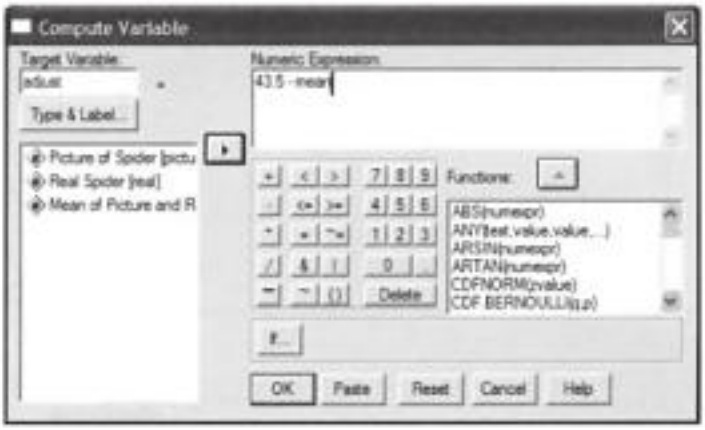


Figura 7.6 Calculando o fator de ajuste.

Essa diferença pode ser usada para ajustar os escores existentes de cada participante. Primeiro, precisamos ajustar os escores na condição **picture of spider** (figura da aranha). Mais uma vez, podemos usar o comando *compute* (calcular) para fazer o ajuste. Ative a caixa de diálogo *compute* da mesma forma que antes e, então, dê o nome para a nova variável de **picture2** (você pode, depois, clicar em Tipo & Rótulo) (Tipo e Rótulo) e dar a essa variável um nome tal como "valores ajustados para a condição figura"). Iremos acrescentar o escore de cada participante na condição **picture of spider** para o seu valor ajustado. Selecione a variável **picture** (figura) e a transfira para a área de comando clicando em ▶. Depois, clique em ▼, selecione a variável **adjust** (ajuste) e a transfira para a área de comando clicando em ▶. A caixa de diálogo completa é mostrada na Figura 7.7. Faça o mesmo com a variável **real spider**: crie uma variável chamada de **real2** que contém os valores da **real** acrescidos ao valor da coluna **adjust.**

As variáveis **real2** e **picture2** (verdadeira2 e figura2) representam a ansiedade experimentada em cada condição ajustada de maneira a eliminar as diferenças entre os participantes. Se você não acredita em mim, use o comando *compute* (calcular) para criar a variável **mean2** (média2) que é a média entre **real2** e **picture2** (verdadeira2 e figura2) (como fizemos na Seção 7.3.2.1). Você deverá verificar que o valor nessa coluna é o mesmo para cada

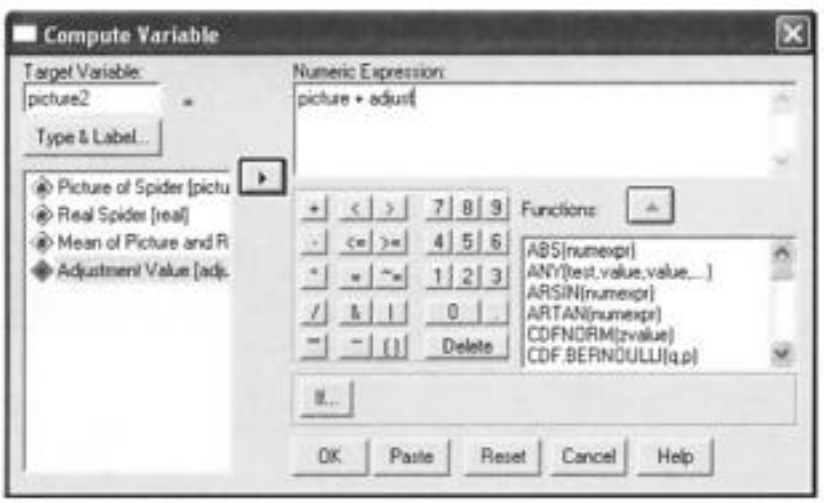


Figura 7.7 Ajustando os valores da condição **picture of spider**.

Material com direitos autorais

participante demonstrando que a variabilidade entre participantes nas médias desapareceu: o valor será 43,50 – a média geral.

7.3.2.5 Traçar o diagrama de barras de erros ②

Traçar o gráfico propriamente dito é semelhante ao cenário entre grupos. Primeiro, acesse a caixa de diálogo principal por meio dos menus **Graphs⇒ErrorBar...**. (Gráficos⇒Barra de Erros...). Depois, nessa caixa de diálogo (veja a Figura 7.1) clique em Summaries of separate variables (Resumo para variáveis separadas) e em Define (Definir). Esse processo irá criar a caixa de diálogo principal. Uma vez que a caixa de diálogo for ativada, selecione as duas variáveis de interesse (queremos ajustar valores, assim, precisamos selecionar **real2** e **picture2** (verdadeira2 e figura2)) e, então, clique em OK (veja a Figura 7.8).

O gráfico de barras de erro resultante é apresentado na Figura 7.9. Você irá notar que o gráfico de barras de erro dos valores ajustados é um pouco diferente do gráfico de barras de erro dos valores não-ajustados (retorne à Figura 7.3). A primeira coisa a ser notada é que as barras de erro não se sobrepõem. Anteriormente falei que quando as barras de erro não se sobrepõem podemos estar confiantes de que nossas amostras não vieram da mesma população (e, assim, nossa manipulação experimental foi um sucesso). Portanto, a versão do gráfico de medidas repetidas desses dados indica que aranhas verdadeiras podem evocar maiores graus de ansiedade (porque a média é alta) do que apenas uma a figura de uma aranha. Esse exemplo ilustra que conclusões diferentes podem ser tiradas se os dados foram coletados usando os mesmos ou diferentes participantes (falarei mais sobre isso na Seção 7.7).

7.4 TESTANDO DIFERENÇAS ENTRE MÉDIAS: O TESTE *T* ①

A forma mais simples de fazer um experimento é manipular uma só variável de duas maneiras apenas e medir somente uma saída. A manipulação da variável independente em geral envolve ter uma condição experimental e um grupo-controle (veja Field e Hole, 2003). Alguns exemplos de tipos de projetos são:

- O filme *Pânico 2* é mais assustador do que o *Pânico* original? Podemos medir os bati-

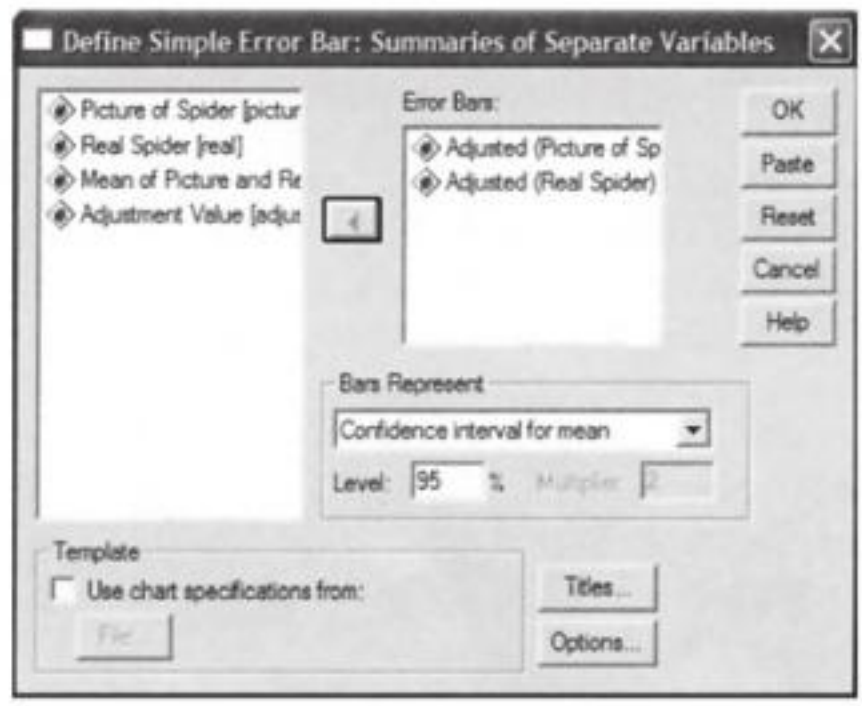


Figura 7.8 Caixa de diálogo para traçar um gráfico de barras de erro de uma variável de medidas repetidas.

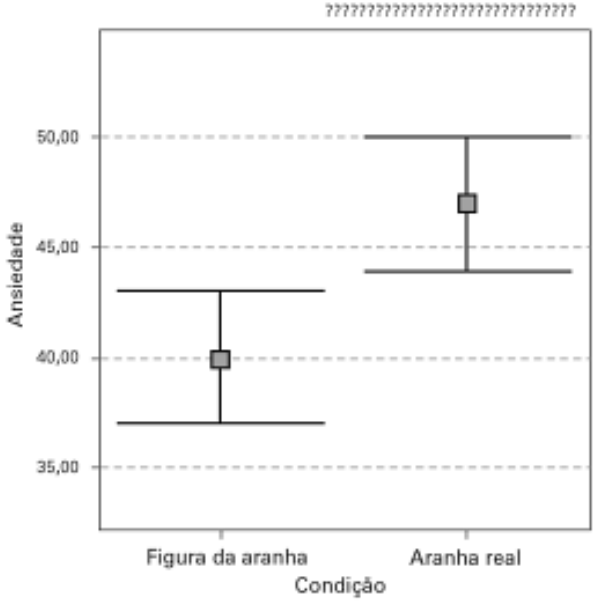


Figura 7.9 Gráfico da barra de erro dos valores ajustados de **spiderRM.sav**.

mentos cardíacos (que indicam ansiedade) durante ambos os filmes e compará-los.
- Escutar música enquanto você trabalha melhora o seu desempenho? Você poderia pedir para algumas pessoas escreverem um artigo (ou livro!) escutando sua música favorita e, então, escreverem outro artigo trabalhando em silêncio (isso é um grupo-controle). Depois, você poderia comparar as notas dos artigos!
- Escutar a música favorita do Andy melhora o seu desempenho? Você poderia repetir a estratégia acima, mas, em vez de deixar as pessoas trabalharem ouvindo sua música favorita, você poderia colocar algumas das minhas músicas preferidas (listadas nos agradecimentos) e ver a qualidade do seu trabalho despencar!

O teste *t* foi projetado para analisar esses cenários. É claro, há projetos experimentais mais complexos e os veremos nos capítulos seguintes. Existem dois tipos de testes *t* e qual você usará irá depender de como a variável independente foi manipulada, se com os mesmos ou diferentes participantes:

- **O teste *t* para amostras independentes:** Esse teste é usado quando existem duas condições experimentais e diferentes participantes foram designados para cada condição (isso às vezes é chamado de teste *t* de *medidas independentes* ou *amostras independentes*).
- **O teste *t* para amostras dependentes:** Esse teste é usado quando existem duas condições experimentais e os mesmos participantes tomaram parte em ambas as condições (esse teste, às vezes, é referido como o teste *t para amostras emparelhadas*).

7.4.1 Fundamentação do teste *t* ①

Os dois testes tem uma fundamentação semelhante:

- Duas amostras de dados são coletadas e a média das amostras é calculada. Essas médias podem diferir pouco ou muito.
- Se as amostras vêm da mesma população, esperamos que suas médias sejam praticamente iguais (veja a Seção 1.6). Embora seja possível que as médias difiram somen-

Material com direitos autorais

te pelo acaso, grandes diferenças entre as médias amostrais não devem ocorrer com frequência. Sob o que é conhecido como *hipótese nula,* presumimos que a manipulação experimental não tenha efeito nos participantes: portanto, esperamos que as médias das amostras sejam similares.

- Comparamos a diferença entre as médias das amostras que coletamos com a diferença entre as médias das amostras que esperamos obter somente pelo acaso. Usamos o erro padrão (veja a Seção 1.6) como uma medida da variabilidade entre as médias das amostras. Se o erro padrão é pequeno, esperamos que a maioria das amostras tenham médias similares. Quando o erro padrão é grande, esperamos obter grandes diferenças nas médias das amostras somente pelo acaso. Se a diferença entre as amostras que coletamos for maior do que esperamos no erro padrão, podemos presumir uma de duas coisas:
  a) Que as médias das amostras na nossa população flutuam muito somente por acaso e temos, por acaso, coletado duas amostras que são atípicas da população de onde foram retiradas.
  b) As duas amostras vêm de populações diferentes, mas são típicas das suas respectivas populações originais. Nesse cenário, a diferença entre amostras representa uma diferença genuína entre as amostras (e, assim, a hipótese nula é incorreta).
- À medida que a diferença observada entre as médias das amostras aumenta mais confiante ficamos de que a segunda explicação seja a correta (isto é, que a hipótese nula deva ser rejeitada). Se a hipótese nula é incorreta, aceitamos o que é denominado de *hipótese experimental* (também chamada de *hipótese alternativa* ou algo semelhante), isto é, que as médias das duas amostras diferem por causa da manipulação diferenciada que foi imposta a cada uma das amostras.

Resumindo, calculamos o teste *t* usando a equação (7.1), mas a forma exata que essa equação tomará vai depender de como os dados foram coletados (isto é, se os mesmos ou diferentes participantes foram usados em cada condição experimental).

$$t = \frac{\begin{array}{c}\text{diferença} \\ \text{observada} \\ \text{entre as} \\ \text{médias} \\ \text{amostrais}\end{array} - \begin{array}{c}\text{diferença esperada} \\ \text{entre as médias} \\ \text{populacionais} \\ \text{(se a hipótese nula} \\ \text{for verdadeira)}\end{array}}{\begin{array}{c}\text{estimativa do erro padrão} \\ \text{da diferença entre as} \\ \text{médias de duas amostras}\end{array}} \qquad (7.1)$$

7.4.2 Suposições do teste *t* ①

Tanto o *teste t independente* como o *dependente* são *testes paramétricos* baseados na distribuição normal (veja o Capítulo 3). Portanto, é assumido que:

- Os dados são de populações normalmente distribuídas.
- Os dados são medidos pelo menos em um nível intervalar.

O teste *t* independente, que é usado para testar diferentes grupos de pessoas, também assume que:

- As variâncias populacionais são iguais (*homogeneidade da variância).*
- Os escores são independentes (porque eles vêm de diferentes pessoas).

Essas hipóteses foram explicadas em detalhes no Capítulo 3, no qual enfatizei a necessidade de verificar essas suposições antes de realizar seu teste estatístico. Não vou falar sobre isso novamente, mas se você ignorou meu conselho e não verificou essas suposições, então precisa fazê-lo agora! O SPSS incorpora, também, procedimentos do teste *t* (por exemplo, o teste de Levene – veja a Seção 3.6 – pode ser feito ao mesmo tempo em que o teste *t*). Vamos agora analisar cada um dos testes *t*.

7.5 O TESTE *t* DEPENDENTE ①

Agora que sabemos como fazer gráficos de médias, podemos continuar a analisar se diferenças entre as médias dos grupos são esta-

Material com direitos autorais

tisticamente significativas. Se ficarmos com os nossos dados de medidas repetidas, podemos olhar para o teste *t* dependente, ou teste *t* de amostras emparelhadas. O teste *t* dependente é fácil de calcular. Usamos a versão numérica da equação (7.1):

$$t = \frac{\bar{D} - \mu_D}{S_D / \sqrt{N}} \qquad (7.2)$$

A equação (7.2) compara a diferença média entre nossas amostras ($\bar{D}$) com a diferença que devemos esperar encontrar entre as médias populacionais (μ_D) e, então, leva em conta o erro padrão das diferenças ($S_D/\sqrt{N}$). Se a hipótese nula é verdadeira, esperamos que não haja diferenças entre as médias populacionais (isto é, $\mu_D = 0$).

7.5.1 Distribuições amostrais e o erro padrão ①

Na equação (7.1), me referi ao denominador da fração anterior como o erro padrão das diferenças. O erro padrão foi apresentado na Seção 1.6, e é simplesmente o desvio padrão da *distribuição amostral*. Dê uma olhada nessa seção agora para refrescar sua memória sobre distribuições amostrais e erro padrão. As distribuições amostrais têm muitas propriedades importantes. Para começar, se a população é distribuída normalmente, então o mesmo ocorre com a distribuição amostral; na verdade, se as amostras contêm mais do que 50 escores, a distribuição amostral será sempre distribuída normalmente. A média da distribuição amostral é igual à média da população – assim, se você calculou a média de todas as médias das amostras, o valor que você obtém deve ser o mesmo da média da população. Essa propriedade faz sentido porque se uma amostra é representativa da população, você deve esperar que sua média seja igual à da população. Entretanto, algumas vezes amostras não são representativas e sua média difere da média da população. Em geral, entretanto, a média da amostra estará bem próxima da média da população e raramente a média da amostra será substancialmente diferente daquela da população. Finalmente, outra propriedade da distribuição amostral é que seu desvio padrão é igual ao desvio padrão da população dividido pela raiz quadrada do número de observações na amostra. Como mencionei anteriormente, esse desvio padrão é conhecido como o erro padrão.

Podemos estender essa ideia olhando para as *diferenças* entre as médias das amostras. Se você fosse tirar vários pares de amostras da população e calcular suas médias, poderia, também, calcular a diferença entre suas médias. Mencionei anteriormente que em geral as médias das amostras serão muito similares à média da população, portanto, a maioria das amostras terá médias muito similares. Desse modo, na maior parte do tempo, a diferença entre as médias das amostras de uma mesma população será zero ou próxima de zero. Entretanto, algumas vezes uma ou ambas as amostras podem ter um média muito afastada da média da população e, assim, é possível obter uma grande diferença entre as médias das amostras somente por acaso. Entretanto, isso acontece com menos frequência.

Na verdade, se traçar essas diferenças entre as médias das amostras como um histograma você terá, novamente, uma distribuição amostral com todas as propriedades previamente descritas. O desvio padrão dessa distribuição amostral é chamado de *erro padrão das diferenças*. Um erro padrão pequeno indica que a maioria dos pares das amostras de uma população terá médias muito parecidas (isto é, a diferença entre as médias da amostra será frequentemente muito pequena). Um erro padrão grande indica que as médias das amostras podem se desviar muito da média da população e, assim, diferenças entre pares de amostras podem ser bem grandes somente por acaso.

7.5.2 A equação do teste *t* dependente explicada ①

Em um experimento, o escore de uma pessoa na condição 1 será diferente do seu escore na condição 2, e essa diferença pode ser muito grande ou muito pequena. Se calculássemos as diferenças entre o escore de cada pessoa em cada condição e somássemos essas diferenças,

Material com direitos autorais

teríamos o total das diferenças. Se, depois, dividíssemos essa diferença pelo número de participantes, teríamos a diferença média (ou seja, quanto, em média, o escore de uma pessoa diferiu na condição 1 comparado com a condição 2). Essa diferença média é ($\bar{D}$) na equação (7.2) e é um indicador da variação sistemática nos dados (isto é, ele representa o efeito experimental). Se tivéssemos tirado duas amostras aleatórias da população (e nada feito com essas amostras), as médias poderiam ser diferentes somente por acaso. Como já vimos, o grau no qual duas médias da amostra provavelmente diferem é determinado pelo tamanho do erro padrão das diferenças. Precisamos ter certeza de que a diferença observada entre nossas amostras é devido à manipulação experimental (e não resultado do acaso). Saber somente a diferença média não é útil porque ela depende da escala da medida, portanto, padronizamos o valor. Uma maneira de padronizar a soma da diferença dos grupos seria dividi-la pelo desvio padrão da amostra dessas diferenças (veja a Seção 1.4.1). O desvio padrão representa a média dos desvios a partir da média, assim, o desvio padrão das diferenças entre condições representa o desvio médio da diferença das médias. Como tal, o desvio padrão é uma medida de quanta variação existe entre a diferença dos escores dos participantes. Desse modo, o desvio padrão dessas diferenças representa a variação *não-sistemática* no experimento.

Esclarecendo, imagine que um alienígena (quando eu estava paralisado de medo de uma aranha) me clonou 12 vezes (que os céus não permitam que existam 12 iguais a mim no planeta!). Todos os meus clones seriam idênticos a mim e se comportariam do mesmo modo que eu. Portanto, todos nós estaríamos amedrontados com a figura da aranha e teríamos um escore de ansiedade de 40. Ao ver aranhas verdadeiras, todos molharíamos nossa roupa de baixo e teríamos escores de ansiedade de 50. Lembre-se de que somos clones, portanto, temos um comportamento idêntico. Se você calcular a diferença entre os escores da nossa ansiedade para a figura e a aranha verdadeira, a diferença será de 10 pontos para cada um de nós. Se você calcular o desvio padrão dessas diferenças o resultado será 0 (porque todos nós temos os mesmos escores, não há variação). Como nossos participantes são os mesmos, o desvio padrão é 0; isto é, não há variação não-sistemática. Em outras palavras, todas as diferenças em ansiedade resultam da observação de aranhas verdadeiras em vez de figuras de aranhas e nenhuma das diferenças pode ser explicada por outros fatores (como diferenças individuais).

Embora dividir pelo desvio padrão seja útil para padronizar a diferença média entre condições, de fato, estamos interessados em saber como a diferença entre as médias das amostras se compara com a que esperaríamos obter se não tivéssemos feito uma manipulação experimental. Podemos usar as propriedades da distribuição amostral: em vez de dividir a diferença média entre condições pelo desvio padrão das diferenças, podemos dividi-la pelo erro padrão das diferenças. Dividir pelo erro padrão não somente padroniza a diferença média entre condições, mas também informa como a diferença entre as duas médias das amostras se compara em magnitude ao que se esperaria somente por acaso. Se o erro padrão é grande, grandes diferenças entre amostras são mais comuns (porque a distribuição entre as diferenças é mais espalhada). Já se o erro padrão é pequeno, grandes diferenças entre as médias das amostras são incomuns (porque a distribuição está mais concentrada em torno do zero). Portanto, se a diferença média entre nossas amostras é grande e o erro padrão das diferenças é pequeno, podemos estar confiantes de que a diferença que observamos na nossa amostra não é um resultado casual, ou seja, ela deve ter sido provocada pela manipulação experimental.

Em um mundo perfeito, calcularíamos o erro padrão retirando todos os pares possíveis de amostras da população, descobrindo as diferenças entre suas médias e determinando o desvio padrão dessas diferenças. Entretanto,

Material com direitos autorais

na primeira variável com o *mouse*, depois na segunda e note que os nomes das variáveis aparecem na caixa denominada de *Current Selections* (Seleções atuais)) e as transfira para a caixa denominada *Paired Variables* (Variáveis Pareadas) clicando em ▶. Se você quiser fazer vários testes *t*, pode selecionar outro par de variáveis, transferi-las para a lista de variáveis, depois selecionar outro par, e assim por diante. Nesse caso, queremos executar somente um teste. Se você clicar em Options... (opções), aparece outra caixa de diálogo que oferece a opção de mudar a amplitude do intervalo de confiança que é calculado. A configuração padrão é para um intervalo de confiança de 95% e isso é adequado. Entretanto, se você quisesse ser mais rigoroso sobre sua análise, poderia ter escolhido um intervalo de confiança de 99%, mas assim você correria um risco maior de não conseguir detectar um efeito genuíno (um erro do Tipo II). Para executar a análise, clique em OK.

7.5.4 Saídas do teste *t* dependente ①

A saída resultante produz três tabelas. A saída do SPSS 7.1 mostra uma tabela de estatísticas resumo para as duas condições experimentais. Para cada condição, sabemos a média, o número de participantes (N) e o desvio padrão da amostra. A coluna final informa o erro padrão (veja a Seção 7.5.1), que é o desvio padrão da amostra dividido pela raiz quadrada do tamanho da amostra ($EP = s/\sqrt{N}$), assim, para a condição da figura $EP = 9{,}2932/\sqrt{12} = 9{,}2932/3{,}4641 = 2{,}68$.

A Saída do SPSS 7.1 também mostra a correlação de Pearson entre as duas condições. Quando medidas repetidas são usadas, é possível que as condições experimentais se correlacionem (porque os dados em cada condição vêm das mesmas pessoas e, assim, pode haver consistência em suas respostas). O SPSS dá o valor do r de Pearson e a significância bilateral (veja o Capítulo 4). Para esses dados, as condições experimentais produzem um coeficiente de correlação bem grande ($r = 0{,}545$), mas não estão correlacionadas de forma significativa porque $p > 0{,}05$.

A Saída do SPSS 7.2 mostra a tabela mais importante: aquela que nos diz se a diferença entre as médias das duas condições foi grande o suficiente para *não* ser um resultado ao acaso. Primeiro a tabela fornece a diferença da média entre escores (esse valor – $\bar{D}$ na equação (7.2) – é a diferença entre os escores médios de cada condição: $40 - 47 = -7$). A tabela também apresenta o desvio padrão das diferenças entre as médias e, mais importante, o erro padrão das diferenças entre os escores dos participantes em cada condição (veja a Seção 7.5.1). A estatística teste *t* é calculada dividindo-se a média das diferenças pelo erro padrão das diferenças (veja a equação (7.2): $t = -7/2{,}8311 = -2.47$). O tamanho do *t* é comparado contra valores conhecidos com base nos graus de liberdade. Quando os mesmos participantes foram usados, os graus de liberdade são o tamanho da amostra menos 1 ($gl = N - 1 = 11$). O SPSS usa os graus de liberdade para calcular a probabilidade exata de que um valor de *t* tão grande quanto o obtido possa ocor-

Saída do SPSS 7.1

Paired Sample Statistics
(Estatísticas Amostrais Pareadas)

		Mean (Média)	*N*	*Std. Deviation* (Desvio Padrão)	*Std. Error Mean* (Erro Padrão da Média)
Pair 1 (Par 1)	*Picture of Spider* (Figura da Aranha)	40.0000	12	9.2932	2.6827
	Real Spider (Aranha Verdadeira)	47.0000	12	11.0289	3.1838

Paired Sample Correlations
(Correlações Amostrais Pareadas)

		N	*Correlation* (Correlação)	*Sig.*
Pair 1 (Par 1)	*Picture of Spider* (Figura da Aranha) *Real Spider* (Aranha Verdadeira)	12	0.545	0.067

Material com direitos autorais

Saída do SPSS 7.2

Paired Samples Test **(Testes Amostrais Pareados)**

		Paired Differences (Diferenças pareadas)					*t*	*df* (gl)	*Sig. (2-tailed)* (Sig. bilateral)
		Mean (Média)	*Std. Deviation* (Desvio Padrão)	*Std. Error Mean* (Erro Padrão da Média)	*95% confidence interval of the differences* (Intervalo de confiança de 95% das diferenças)				
					Lower (Inferior)	*Upper* (Superior)			
Pair 1 (Par 1)	*Picture of Spider* (Figura da Aranha) *Real Spider* (Aranha Verdadeira)	–7.0000	9.8072	2.8311	–13.2312	–0.7688	–2.473	11	0.031

rer por acaso. Essa probabilidade está na coluna chamada de Sig. (Significância). Por padrão, o SPSS fornece a probabilidade bilateral, que é a probabilidade quando nenhuma previsão foi feita sobre a direção das diferenças entre os grupos. Se uma previsão específica foi feita (por exemplo, podemos prever que a ansiedade será maior quando uma aranha verdadeira é usada), então a probabilidade unilateral deve ser informada, e esse valor é obtido dividindo-se a probabilidade bilateral por 2. A probabilidade bilateral para os dados da aranha é muito baixa ($p = 0{,}031$); ela nos diz que há somente uma chance de 3,1% de que um valor de *t* desse tamanho possa ter ocorrido devido ao acaso. Como cientistas sociais, aceitamos como estatisticamente significativo tudo o que tem menos de 5% de chance de ocorrer por acaso. O fato de o valor de *t* ser um número negativo nos diz que a condição 1 (a condição **picture of spider** (figura da aranha)) tem uma média menor do que a segunda (a condição **real spider** (aranha derdadeira)) e, assim, a aranha verdadeira provoca uma ansiedade maior do que a da figura. Quando apresentamos um teste *t*, sempre incluímos os graus de liberdade (nesse caso, 11), o valor da estatística *t* e o nível no qual esse valor é significativo. Portanto, podemos concluir que a exposição à aranha verdadeira causou mais ansiedade em aracnofóbicos do que a exposição a uma figura ($t(11) = -2{,}47$, $p < 0{,}05$). Esse resultado foi previsto pelo gráfico de barra de erros apresentado na Figura 7.9.

A última informação que essa saída nos fornece é um intervalo de confiança de 95% para a diferença média. Imagine que tomamos 100 amostras de uma população de diferentes escores, calculamos suas médias ($\bar{D}$) e um intervalo de confiança para essa média. Em 95 das amostras, o intervalo de confiança construído contém o valor real da diferença média. O intervalo de confiança fornece os limites dentro dos quais é mais provável que a diferença real média esteja[3]. Assim, assumindo que o intervalo de confiança da amostra é um dos 95 dos 100 que contém o valor da população, podemos dizer que a diferença média verdadeira está entre $-13{,}23$ e $-0{,}77$. A importância desse intervalo é que ele não contém o zero (isto é, ambos os limites são negativos); isso nos diz que o valor verdadeiro da diferença média dificilmente será zero. Se comparássemos pares de amostras aleatórias de uma população, esperaríamos que a maioria das diferenças entre as médias das

[3] Vimos na Seção 1.6.2 que esses intervalos representam o valor de dois (1,96, precisamente) erros padrão para cada lado da média da distribuição amostral. Para esses dados em que a diferença média foi -7 e o erro padrão foi de 2,8311, esses limites serão de $-7 \pm (1{,}96 \times 2{,}8311)$. Contudo, em virtude de estarmos utilizando a distribuição *t* e não a distribuição normal usaremos um valor crítico da distribuição *t* para calcular os intervalos de confiança. Esse valor é (com 11 graus de liberdade como nesse exemplo) 2,201 (bilateral), que fornece um intervalo de $-7 \pm (2{,}201 \times 2{,}8311)$.

Material com direitos autorais

amostras fosse zero. Esse intervalo nos diz que, com base nas nossas duas amostras, é pouco provável que o valor verdadeiro entre as médias seja zero. Portanto, podemos estar confiantes de que nossas duas amostras não representam amostras aleatórias da mesma população. Em vez disso, elas representam amostras de populações diferentes induzidas pela manipulação experimental.

7.5.5 Calculando o tamanho de efeito ②

Embora nosso t estatístico seja estatisticamente significativo, isso não quer dizer que nosso efeito seja importante em termos práticos. Para descobrir se o efeito é importante, precisamos usar o que sabemos sobre o tamanho de efeito (veja a Seção 1.8.4). Ficarei com o tamanho de efeito r porque ele é amplamente entendido, frequentemente utilizado e sim, admito, eu realmente gosto dele! Converter um valor t em um valor r é muito fácil; usamos a seguinte equação (de Rosenthal, 1991, p.19 e Rosnow e Rosenthal, 2005, p.328):

$$r = \sqrt{\frac{t^2}{t^2 + gl}}$$

Sabemos os valores de t e do gl da saída do SPSS, assim, podemos calcular r da seguinte maneira:

$$\begin{aligned} r &= \sqrt{\frac{2{,}473^2}{-2{,}473^2 + 11}} \\ &= \sqrt{\frac{6{,}116}{17{,}116}} \\ &= 0{,}60 \end{aligned}$$

Se você lembrar os nossos pontos de referência para os tamanhos de efeito, isso representa um efeito muito grande (ele está acima de 0,5, o limite para um efeito grande). Portanto, além de estatisticamente significativo, esse efeito também é grande; desse modo, representa uma descoberta significativa.

7.5.6 Relatando o teste *t* dependente ①

Existe uma maneira padrão para relatar qualquer estatística teste: você declara a descoberta com a qual o teste se relaciona, e depois relata a estatística teste, seus graus de liberdade e o valor de probabilidade dessa estatística teste. A Associação Americana de Psicologia, entre outras associações, recomenda que uma estimativa do tamanho de efeito seja rotineiramente relatada. Mesmo que os tamanhos de efeito sejam usados esporadicamente, quero que você tenha bons hábitos; portanto, vamos começar a pensar em tamanhos de efeito agora. Nesse exemplo, a saída do SPSS indica que o valor do t era $-2{,}47$, que os graus de liberdade nos quais isso foi baseado era 11 e que era significativo em $p = 0{,}031$. Também é possível ver as médias para cada grupo. Podemos escrever isso como:

✓ Na média, os participantes experienciaram maior ansiedade com aranhas verdadeiras ($M = 47{,}00$, $EP = 3{,}18$) do que com figuras de aranhas ($M = 40{,}00$, $EP = 2{,}68$, $t(11) = -2{,}43$, $p < 0{,}05$, $r = 0{,}60$).

Observe como relatamos as médias em cada grupo (e erros padrão) no formato padrão. Para a estatística teste, note que usamos um t em itálico a fim de representar o fato de que calculamos uma estatística t, e entre parênteses colocamos os graus de liberdade e apresentamos o valor da estatística teste. A probabilidade pode ser expressa de várias maneiras: geralmente, relatamos pesquisas em um nível padrão de significância (como 0,05), conforme fizemos aqui, mas, às vezes, podemos relatar a significância exata. Finalmente, note que relatei o tamanho de efeito no final – dificilmente você verá isso em artigos publicados, mas deve relatá-lo mesmo assim!

Evite escrever de forma vaga, como:

✘ As pessoas estavam mais assustadas com as aranhas verdadeiras ($t = -2{,}47$)

Mais assustadas em comparação a quê? Onde estão os gl? O resultado foi estatisticamente significativo? O efeito era importante (qual é o tamanho de efeito?)

Material com direitos autorais

Dica da Samanta Ferrinho

- O teste *t* dependente compara duas médias quando essas médias vieram do mesmo grupo de pessoas; por exemplo, se você tem duas condições experimentais e usou os mesmos participantes em cada condição.
- Veja a coluna rotulada de *Sig*. Se o valor for menor do que 0,05, as médias dos dois grupos são significativamente diferentes.
- Veja os valores das médias para informar como os grupos diferem.
- O SPSS fornece somente a significância bilateral; se você quiser a significância unilateral, basta dividir o valor por 2.
- Relate a estatística *t*, os graus de liberdade e o valor da significância. Relate, também, as médias e os seus erros padrão correspondentes (ou faça um gráfico de barras de erro).
- Se você for corajoso, calcule o tamanho de efeito e relate isso, também!

7.6 O TESTE *T* INDEPENDENTE ①

7.6.1 A equação do teste *t* independente explicada ①

O teste *t* independente é usado em situações nas quais existem duas condições experimentais e participantes diferentes foram usados em cada uma. Duas equações diferentes podem ser utilizadas para calcular a estatística *t* dependendo se as amostras contêm ou não um número igual de participantes. Assim como com o teste *t* dependente, podemos calcular a estatística *t* usando uma versão numérica da equação (7.1). Com o teste *t* dependente, podíamos calcular as diferenças entre os pares de escores porque eles vinham dos mesmos participantes e, assim, diferenças individuais entre condições eram eliminadas. Dessa maneira, a diferença no escore refletia somente o efeito da manipulação experimental. Já quando participantes diferentes representam condições diferentes, os pares dos escores irão diferir não somente por causa da manipulação experimental, mas também devido a outras fontes de variação (como diferenças individuais entre a motivação dos participantes, QI, etc.). Se não podemos investigar as diferenças entre as condições em uma base *por participante* (comparando pares de escores como fizemos com o teste *t* dependente), devemos fazer comparações em bases *por condição* (olhando para o efeito geral na condição – veja a equação (7.3)):

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\text{Estimativa do erro padrão}} \quad (7.3)$$

Em vez de observar as diferenças entre pares de escores, agora olhamos para as diferenças entre as médias gerais das duas amostras e as comparamos com as diferenças que esperamos conseguir entre as médias das duas populações de onde as amostras vieram. Se a hipótese nula é verdadeira, as amostras foram retiradas da mesma população. Portanto, sob a hipótese nula $\mu_1 = \mu_2$ e $\mu_1 - \mu_2 = 0$. Assim, a equação fica:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{Estimativa do erro padrão}} \quad (7.4)$$

No teste *t* dependente, dividimos a diferença média entre pares de escores pelo erro padrão dessas diferenças. No teste *t* independente, estamos olhando para as diferenças entre grupos, assim, precisamos dividir pelo desvio padrão das diferenças entre os grupos. Também podemos aplicar a lógica das distribuições amostrais nessa situação. Imagine que pegamos vários pares de amostras – cada par contendo uma amostra de cada uma das duas diferentes populações e comparamos as médias dessas amostras. Do que aprendemos sobre distribuições amostrais, sabemos que a maioria das amostras de uma população terá médias muito parecidas. Portanto, se pegarmos vários pares de amostras (de popu-

Material com direitos autorais

lações diferentes), as diferenças entre as médias das amostras serão semelhantes entre os pares. Entretanto, geralmente a diferença entre as médias de um par de amostras irá se desviar por um pequeno valor, e, muito ocasionalmente, irá desviar por um valor maior. Se pudéssemos traçar uma distribuição amostral das diferenças entre cada par de médias amostrais que pudéssemos obter de duas populações, descobriríamos que ela teria uma distribuição normal com uma média igual à diferença entre as médias das populações ($\mu_1 - \mu_2$). A distribuição amostral nos informaria o quanto podemos esperar que as médias de duas (ou mais) amostras difiram. Como anteriormente, o desvio padrão da distribuição amostral (o erro padrão) indica quão variáveis serão as diferenças entre as médias das amostras apenas como consequência do acaso. Se o desvio padrão é alto, grandes diferenças entre as médias das amostras podem ocorrer por acaso; se é pequeno, somente pequenas diferenças entre as médias das amostras são esperadas. Portanto, faz sentido usar o erro padrão da distribuição amostral para avaliar se a diferença entre as duas médias das amostras é estatisticamente significativa ou simplesmente fruto do acaso. Especificamente, dividimos as diferenças entre as médias das amostras pelo desvio padrão da distribuição amostral.

Mas como obtemos o desvio padrão da distribuição amostral das diferenças entre médias amostrais? Bem, utilizamos a *propriedade da soma das variâncias*, a qual afirma que a variância da diferença entre duas variáveis independentes é igual à soma das suas variâncias (veja Howell, 2002, Capítulo 7). Essa afirmação significa que a variância da distribuição amostral é igual à soma das variâncias das duas populações de onde as amostras foram coletadas. Vimos anteriormente que o erro padrão é o desvio padrão da distribuição amostral de uma população. Podemos usar os desvios padrão amostrais para calcular o erro padrão da distribuição amostral de cada população:

$$\text{EP da distribuição amostral da primeira população} = \frac{s_1}{\sqrt{N^1}}$$

$$\text{EP da distribuição amostral da segunda população} = \frac{s_2}{\sqrt{N^2}}$$

Portanto, lembrando que a variância é simplesmente o desvio padrão ao quadrado, podemos calcular a variância de cada distribuição amostral:

$$\text{Variância da distribuição amostral da primeira população} = \left(\frac{s_1}{\sqrt{N_1}}\right)^2 = \frac{s_1^2}{N_1}$$

$$\text{Variância da distribuição amostral da segunda população} = \left(\frac{s_2}{\sqrt{N_2}}\right)^2 = \frac{s_2^2}{N_2}$$

A lei da soma das variâncias significa que para encontrar a variância da distribuição amostral das diferenças simplesmente somamos as variâncias das distribuições amostrais das duas populações:

$$\text{Variância da distribuição amostral das diferenças} = \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}$$

Para encontrar o erro padrão da distribuição amostral das diferenças, tiramos a raiz quadrada da variância (porque a variância é o desvio padrão ao quadrado):

$$\text{EP da distribuição amostral das diferenças} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

Portanto, a equação (7.4) se torna a equação (7.5):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \qquad (7.5)$$

A equação (7.5) é verdadeira somente quando os tamanhos da amostra são iguais. Em geral, nas ciências sociais não é possível coletar amostras de mesmo tamanho (porque, por exemplo, nem todas as pessoas completam um experimento). Quando queremos comparar dois grupos que contêm números diferentes de participantes, a equação (7.5) não é apropriada. Em vez disso, usamos a va-

Material com direitos autorais

A segunda tabela da saída (Saída do SPSS 7.4) contém a estatística teste principal. Note que há duas filas contendo valores para a estatística teste: um fila é rotulada de *Variâncias iguais assumidas*, e a outra é rotulada de *Variâncias iguais não assumidas*. No Capítulo 3, vimos que os testes paramétricos supõem que as variâncias em grupos-experimentais são aproximadamente iguais. Bem, na verdade existem ajustes que podem ser feitos em situações nas quais as variâncias não são iguais. As filas da tabela relatam se essa suposição foi quebrada ou não. Como sabemos se essa suposição foi quebrada ou não?

Podemos apenas olhar aos valores da variância e ver se eles são similares (por exemplo, sabemos que os desvios padrão dos dois grupos são 9,29 e 11,03; se elevarmos ao quadrado esses valores teremos variâncias de 86,30 e 121,66). Entretanto, essa medida seria muito subjetiva: "Oh, veja, a variância no grupo 1 é somente 3000 vezes maior do que a variância no grupo 2: isso é praticamente igual". Felizmente, existe um teste que pode ser feito para ver se as variâncias são diferentes o suficiente para causar preocupação. O teste de Levene (veja a Seção 3.6) é similar ao teste *t*, pois ele testa a hipótese de que as variâncias nos dois grupos são iguais (isto é, a diferença entre as variâncias é zero). Portanto, se o teste de Levene é significativo em $p < 0,05$, podemos concluir que a hipótese nula está incorreta e que as variâncias são significativamente diferentes – desse modo, a suposição de homogeneidade das variâncias foi violada. Se, entretanto, o teste de Levene for não-significativo (isto é, $p > 0,05$), precisamos aceitar a hipótese nula de que a diferença entre as variâncias é zero – as variâncias são praticamente iguais e a suposição é convincente. Para esses dados, o teste de Levene não é significativo (porque $p = 0,386$, que é maior do que 0,05); assim, devemos ler a estatística teste na linha chamada de *Variâncias iguais assumida.* Se o teste de Levene tivesse sido significativo, teríamos lido a estatística teste na linha chamada de *Variâncias iguais não-assumidas.*

Determinado que a suposição de homogeneidade das variâncias está satisfeita, podemos

Saída do SPSS 7.4

Independent Samples Test **(Testes Amostrais Independentes)**

		Levene's test for Equality of Variances (Teste de Levene de Homogeneidade de Variâncias)		*t-test for Equality of Means* (Teste t de Igualdade de Médias)						
									95% Confidence Intervalo f the Mean (Intervalo de Confiança de 95% para a Média)	
		F	*Sig.*	*t*	*df* (gl)	*Sig. (2-tailed)* Sig. Bilateral	*Mean Difference* (Diferença das Médias)	*Std. Error Difference* (Erro Padrão da Diferença)	*Lower* (Inferior)	*Upper* (Superior)
Anxiety (Ansiedade)	*Equal variances assumed* (Supondo Variâncias Homogêneas)	0.782	0.386	–1.681	22	0.107	7.0000	4.1633	–15.6342	1.6342
	Equal Variances Not Assumed (Supondo Variâncias Não-Homogêneas)			–1.681	21.385	0.107	7.0000	4.1633	–15.6486	1.6486

Material com direitos autorais

executar o teste *t* propriamente dito. Sabemos a diferença das médias ($\overline{X}_1 - \overline{X}_2 = 40 - 47 = -7$) e o erro padrão da distribuição amostral das diferenças é calculado utilizando-se o denominador da equação (7.5):

$$\sqrt{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)} = \sqrt{\left(\frac{9{,}29^2}{12} + \frac{11{,}03^2}{12}\right)}$$
$$= \sqrt{(7{,}19 + 10{,}14)}$$
$$= \sqrt{17{,}33}$$
$$= 4{,}16$$

A estatística *t* é calculada dividindo a diferença média pelo erro padrão da distribuição amostral das diferenças ($t = -7/4{,}16 = -1{,}68$). O valor de *t* é, então, avaliado em relação ao valor de *t* que você pode esperar conseguir ao acaso quando você tem certos graus de liberdade. Para o teste *t* independente, os graus de liberdade são calculados somando os dois tamanhos amostrais e, depois, subtraindo o número de amostras utilizado, nesse caso duas ($df = N_1 + N_2 - 2 = 12 + 12 - 2 = 22$). O SPSS produz o valor exato da significância de *t* e queremos saber se esse valor é menor ou maior do que 0,05. Nesse caso, o valor bilateral de *p* é 0,107, o que é maior do que 0,05; portanto, não houve diferença significativa entre as médias dessas duas amostras. Nos termos do experimento, podemos inferir que aracnofóbicos ficam igualmente ansiosos tanto com a figura da aranha quanto com sua presença verdadeira.

A probabilidade bilateral é utilizada quando uma previsão específica sobre a direção do nosso efeito não foi feita (veja a Seção 1.8.2). Por exemplo, se não soubermos se a aranha verdadeira irá provocar mais ou menos ansiedade, devemos usar o teste bilateral. Entretanto, geralmente em pesquisas podemos fazer previsões específicas sobre qual grupo tem a média maior. Nesse exemplo, provavelmente teríamos previsto que a aranha verdadeira provocaria mais ansiedade do que a figura e que, consequentemente, a média do grupo verdadeira (*real*) seria maior do que a média do grupo figura (*picture*). Nesse caso, podemos usar o teste unilateral (para mais detalhes sobre esse assunto, veja a Seção 1.8.2, ou Rowntree, 1981, Capítulo 7). Alguns estudantes ficam confusos com o fato de que o SPSS produz somente a significância bilateral e não tem uma opção para produzir a significância unilateral. O motivo é simples: não há necessidade de uma opção porque a probabilidade unilateral pode ser averiguada dividindo o valor da significância bilateral por 2. Nesse caso, a probabilidade bilateral foi 0,107, portanto, a probabilidade unilateral é 0,054 (0,107/2). A probabilidade unilateral é maior do que 0,05 (embora por uma pequena margem), desse modo, concluímos que os aracnofóbicos expostos à aranha verdadeira estavam tão ansiosos quanto os aracnofóbicos expostos à figura da mesma aranha. Esse resultado foi previsto pelo gráfico de barra de erros da Figura 7.3.

7.6.4 Calculando o tamanho de efeito ①

Para descobrir se o nosso efeito é importante, podemos usar a mesma equação da Seção 7.5.5 a fim de converter a estatística *t* em um valor de *r*. Sabemos o valor de *t* e do *gl* da saída do SPSS, podemos calcular *r* da seguinte maneira:

$$r = \sqrt{\frac{-1{,}681^2}{-1{,}681^2 + 22}}$$
$$= \sqrt{\frac{2{,}826}{24{,}826}}$$
$$= 0{,}34$$

Lembrando dos nossos pontos de referência para os tamanhos de efeito, esse resultado representa um efeito médio (ele está por volta de 0,3, o limite para um efeito médio). Portanto, mesmo que o efeito não tenha sido significativo, ele ainda representa um efeito substancial. Você também pode notar que o efeito encolheu, o que parece um pouco estranho dado que usamos exatamente os mesmos dados (mas veja a Seção 7.7)!

7.6.5 Relatando o teste *t* independente ①

As regras que relatei para o teste *t* dependente se aplicam quase que totalmente para o teste *t* independente. A saída do SPSS nos in-

Material com direitos autorais

Dica da Samanta Ferrinho

- O teste *t* independente compara duas médias quando elas vieram de diferentes grupos de pessoas; por exemplo, se você tem duas condições experimentais e usou participantes diferentes em cada condição.
- Veja a coluna rotulada de *Teste de Levene para a Igualdade de Variâncias*. Se o valor de *Sig* é menor do que 0,05, a hipótese de homogeneidade da variância foi violada e você deve olhar somente para a linha na tabela rotulada de *Variâncias iguais não presumidas*. Se o valor de *Sig* do teste de Levene é maior do que 0,05, você deve olhar para a linha na tabela rotulada de *Variâncias iguais presumidas*.
- Olhe para a coluna rotulada de *Sig*: se o valor é menor do que 0,05, as médias dos dois grupos são significativamente diferentes.
- Olhe para os valores das médias para saber como os grupos diferem.
- O SPSS fornece somente o valor de significância bilateral; se você quiser a significância unilateral, basta dividir o valor por 2.
- Relate a estatística *t*, os graus de liberdade e o valor de significância. Indique, também, as médias e seus erros padrão correspondentes (ou trace um gráfico de barra de erros).
- Se você for corajoso, calcule o tamanho de efeito e relate-o também!

forma que o valor de *t* era – 1,68, que os graus de liberdade nos quais isso foi baseado eram 22, e que não era significativo para $p < 0,05$. Também conseguimos ver as médias para cada grupo. Podemos escrever isso como:

✓ Em média, os participantes tiveram mais ansiedade com a aranha verdadeira (M = 47,00, EP = 3,18) do que com a figura de uma aranha (M = 40,00, SE = 2,68). Essa diferença não foi significativa $t(22) = -1,68$, $p > 0,05$; entretanto, ela representou um tamanho de efeito médio $r = 0,34$.

Observe que relatamos as médias em cada grupo (e os erros padrão) como anteriormente. Para a estatística teste, tudo é muito parecido com antes, exceto que precisei relatar que *p* era maior do que (>) 0,05 em vez de menor do que (<). Finalmente, note que comentei sobre o tamanho de efeito no final.

7.7 ENTRE GRUPOS OU MEDIDAS REPETIDAS? ①

Os dois exemplos deste capítulo são interessantes porque ilustram a diferença entre os dados coletados usando os mesmos participantes e usando participantes diferentes. Eles utilizam os mesmos escores em cada condição. Os dois exemplos deste capítulo utilizam os mesmos escores nas duas condições. Quando os dados foram analisados como se tivessem vindo dos mesmos participantes, o resultado foi uma diferença significativa entre as médias, mas quando eles foram analisados como se tivessem vindo de participantes diferentes, não houve diferença significativa entre as médias dos grupos. Isso pode parecer uma descoberta desconcertante – afinal, os números eram idênticos em ambos os exemplos. O que isso ilustra é o *poder* relativo do delineamento das medidas repetidas. Quando os mesmos participantes são usados através das condições, a variância não-sistemática (geralmente chamada de variância do erro) é reduzida drasticamente, facilitando a detecção de qualquer variância sistemática. Muitos podem pensar que a maneira de coletar é irrelevante, mas espero ter provado que ela pode fazer a diferença entre detectar uma diferença ou não. De fato, pesquisadores têm conduzido estudos usando os mesmos participantes em condições experimentais, repetido o estudo utilizando participantes diferentes em condições expe-

Material com direitos autorais

rimentais e, então, usado o método de coleta de dados como uma variável independente na análise. Em geral, eles têm encontrado que o método de coleta de dados interage de forma significativa com os resultados encontrados (veja Erlebacher, 1977).

7.8 O TESTE *t* COMO UM MODELO LINEAR GENERALIZADO ②

Pode parecer estranho eu ter escolhido representar o tamanho de efeito para os meus testes *t* usando *r*, o coeficiente de correlação. Você pode estar pensando: "mas correlações mostram relacionamentos, não diferenças entre médias". Eu costumava pensar assim também, até ler um artigo fantástico de Cohen (1968), o qual me fez entender o que eu estava perdendo: o complexo e espinhoso mundo da estatística subitamente ficou muito mais fácil e simples. Bem, nem tão fácil e simples assim, mas era um bom artigo! O que vou dizer poderá não fazer sentido algum para você ou poderá ajudá-lo a perceber que todos os procedimentos estatísticos são basicamente os mesmos: eles são versões mais ou menos elaboradas do coeficiente de correlação!

No Capítulo 5, vimos que o teste *t* foi usado para testar se o coeficiente de regressão de uma previsão era igual a 0. O projeto experimental para o qual o teste *t* dependente é usado pode ser conceitualizado como uma equação de regressão (afinal, existe uma variável independente – previsor – e uma variável dependente – saída). Se quisermos prever nossa saída, podemos usar a equação geral que mencionei em vários locais:

Se quisermos usar um modelo linear, vimos que essa equação geral se torna a equação (5.2), na qual o modelo é definido pela inclinação e intersecção da reta.

$$\text{Saída}_i = (\text{Modelo}_i) + \text{erro}_i$$

A equação (7.6) é muito similar, em que A_i é a variável dependente (saída), b_0 é o intercepto, b_1 é a ponderação do previsor e G_i é a variável independente (previsor). Incluí a mesma equação, mas com algumas das letras trocadas em relação ao que elas representam no experimento da aranha (assim, A = **ansiedade**, G = **grupo**). Quando executamos um experimento com duas condições, a variável independente tem somente dois valores (grupo 1 e grupo 2). Esses grupos podem ser codificados de várias maneiras (no exemplo da aranha, codificamos o grupo 1 com o valor de 0 e o grupo 2 com o valor de 1). Essa variável codificada é chamada de *dummy (variável auxiliar)* e os valores dessa variável representam grupos de pessoas. Os leitores mais corajosos encontraram esta codificação na Seção 5.10:

$$\begin{aligned} A_i &= b_0 + b_1 G_i + \varepsilon_i \\ \text{Ansiedade}_i &= b_0 + b_1 Grupo_i + \varepsilon_i \end{aligned} \qquad (7.6)$$

Usando o exemplo da aranha, sabemos que a média da **ansiedade** do grupo da figura é de 40 e que a variável **grupo** é igual a 0 para essa condição. Veja o que acontece quando a variável **grupo** é igual a 0 (a condição figura); a equação (7.6) se torna (se ignorarmos o termo residual):

$$\begin{aligned} 40 &= b_0 + (b_1 \times 0) \\ b_0 &= 40 \end{aligned}$$

Portanto, b_0 (o intercepto) é igual à média do grupo da figura (isto é, à média do grupo codificado de 0). Agora, vamos ver o que acontece quando a variável **grupo** é igual a 1. Essa condição, a da aranha verdadeira, teve uma média da **ansiedade** ($\overline{X}_{\text{Verdadeira}}$) de 47. Lembrando que acabamos de descobrir que b_0 é igual à média do grupo da figura ($\overline{X}_{\text{Figura}}$), a equação (7.6) se torna:

$$\begin{aligned} 47 &= b_0 + (b_1 \times 1) \\ 47 &= 40 + b_1 \\ b_1 &= 47 - 40 \\ b_1 &= \overline{X}_{\text{Verdadeira}} - \overline{X}_{\text{Figura}} \end{aligned}$$

Portanto, b_1 representa a diferença entre as médias dos grupos. Como tal, podemos representar um experimento de dois grupos como uma equação de regressão na qual o coeficiente da variável independente (b_1) é igual à diferença entre as médias dos grupo e o intercepto (b_0) é igual à média do grupo codificado como 0. Na regressão, o teste *t* é usado para averi-

Material com direitos autorais

guar se o coeficiente de regressão (b_1) é igual a 0, e quando conduzimos um teste *t* com dados em grupos, testamos se a diferença entre as médias dos grupos é igual a 0.

Para provar que não estou inventando isso, use os dados em ***spiderBG.sav*** e execute uma regressão linear simples utilizando **grupo** como previsor e **ansiedade** como saída. **Grupo** é codificado usando esquema de código de 0 e 1 e, assim, representa a *variável auxiliar* descrita acima. A saída do SPSS resultante deve conter a tabela do resumo da regressão mostrada na Saída 7.5 do SPSS. Primeiro note o valor da constante (b_0): é 40, o mesmo da média da categoria base (o grupo da figura). Agora observe que o valor do coeficiente de regressão b_1 é 7, que é a diferença entre as médias dos dois grupos (47 – 40 = 7). Finalmente, se a estatística *t*, que testa se b_1 é significativamente diferente de 0, é a mesma para o teste *t* independente (veja a Saída 7.4 do SPSS), portanto, é o valor da significância.[4]

Esta seção demonstrou que experimentos podem ser representados em termos de modelos lineares; esse conceito é essencial para o entendimento dos capítulos seguintes sobre modelos lineares generalizados.

[4] Na verdade, o valor da estatística *t* é o mesmo, mas é positivo em vez de negativo. Lembre-se da discussão da correlação ponto bisserial na Seção 4.5.6: quando você correlaciona uma variável dicotômica, a direção do coeficiente de correlação depende inteiramente de que casos são designados a quais grupos. Portanto, a direção da estatística *t* aqui é influenciada pelo grupo que selecionamos como a categoria base (a categoria codificada de 0).

7.9 O QUE FAZER SE OS DADOS NÃO FOREM NORMALMENTE DISTRIBUÍDOS? ②

Neste capítulo, vimos que é possível fazer ajustes no teste *t* quando a hipótese de homogeneidade da variância é violada; mas e quando você tiver dados não normalmente distribuídos? Uma possibilidade é tentar corrigir a distribuição usando uma transformação (veja o Capítulo 3), mas nem sempre funciona. Outra solução útil é usar um *teste não-paramétrico*. Esse tipo de teste tem menos hipóteses do que seus equivalentes paramétricos e, assim, é útil quando seus dados violam as hipóteses dos dados paramétricos descritos na Seção 3.2.1. Alguns desses testes estão descritos no Capítulo 13. O equivalente não-paramétrico do *teste t dependente* é chamado de *teste de postos com sinais de Wilcoxon* (Seção 13.3) e o teste *t* independente tem dois equivalentes não-paramétricos (ambos muito similares) chamados de *teste da somas dos postos de Wilcoxon* e *teste de Mann-Whitney* (Seção 13.2). Recomendo a leitura dessas seções antes de continuar.

7.10 O QUE DESCOBRIMOS SOBRE ESTATÍSTICA? ①

Abordamos muitas coisas neste capítulo; o que começou como "um capítulo breve sobre comparação de médias" acabou imenso. Essas tendências perfeccionistas são uma cruz que tenho que carregar!

Começamos o capítulo analisando as diferentes maneiras como os dados podem ser coletados e por que pode ser interessante com-

Saída 7.5 do SPSS Análise de Regressão entre grupos para os dados das aranhas

Coefficients[a] **(Coeficientes)**

Model (Modelo)		*Unstandardized Coefficients* (Coeficientes Não-Padronizados)		*Standardized Coefficients* (Coeficientes Padronizados)	*t*	*Sig.*
		B	*Std. Error* (Erro Padrão)	*Beta*		
1	*Constant* (Constante)	40.000	2.944		13.587	0.000
	Condition (Condição)	7.000	4.163	0.337	1.681	0.107

a. Variável dependente: ansiedade.

Material com direitos autorais

8

COMPARANDO VÁRIAS MÉDIAS: ANOVA (MLG 1)

8.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ①

O capítulo anterior mostrou que podemos analisar as diferenças entre duas condições ou grupos de pessoas. Os testes explorados naquele capítulo são recursos úteis para a pesquisa em ciências sociais, contudo, eles estão limitados a situações em que existem apenas dois níveis da variável independente (isto é, dois grupos-experimentais). É comum a execução de experimentos em que existem três, quatro ou mesmo cinco níveis da variável independente; nesses casos, os testes vistos no Capítulo 7 não são apropriados. No lugar deles, uma técnica denominada **análise de variância** (ou ANOVA para os íntimos) deve ser utilizada. A vantagem da ANOVA é que ela pode ser utilizada para analisar situações nas quais existem diversas variáveis independentes. Nessas situações, a ANOVA informa como essas variáveis independentes interagem umas com as outras e que efeitos essas interações apresentam sobre a variável dependente. Este capítulo começa explicando a teoria da ANOVA quando diferentes participantes são usados (**ANOVA independente**). Depois iremos aprender a executar uma análise de variância no SPSS e a interpretar os resultados.

8.2 A TEORIA POR TRÁS DA ANOVA ②

8.2.1 Taxas de erros infladas ②

Antes de explicar como a ANOVA funciona, vale a pena mencionar por que não executamos testes *t* para comparar todas as combinações de grupos que estão sendo testados. Imagine uma situação em que existem três condições experimentais e estamos interessados nas diferenças entre esses três grupos. Se executássemos testes *t* para cada par de grupos, seria necessário executar três testes: um para comparar os grupos 1 e 2, outro para comparar os grupos 1 e 3 e um para comparar os grupos 2 e 3. Se cada um desses testes *t* utilizasse uma significância de 5%, a probabilidade de falsamente rejeitar a hipótese nula (conhecido como erro do Tipo I) para cada teste seria de somente 5%. Dessa forma, a probabilidade de não cometer o erro do Tipo I seria 0,95 (95%) para cada teste. Se assumirmos que cada teste é independente (ou seja, podemos multiplicar as probabilidades), a probabilidade total de não cometermos erro do Tipo I seria $(0,95)^3 = 0,857$, porque a probabilidade de não cometermos o erro do Tipo I em cada teste seria 0,95 e existem três testes. Dado

Material com direitos autorais

que a probabilidade de não cometer o erro do Tipo I é 0,857, é possível calcular a probabilidade de cometer pelo menos um erro do Tipo I subtraindo esse resultado de 1 (lembre que a probabilidade máxima de algum evento ocorrer é 1). Assim, a probabilidade de pelo menos um erro do Tipo I ser cometido é 1 – 0,857 = 0,143 ou 14,3%. Dessa forma, nesse conjunto de testes, a probabilidade de cometer erro do Tipo I aumentou de 5% para 14,3%, um valor bem maior que o critério aceito por cientistas sociais. Essa taxa de erro envolvendo um conjunto de testes conduzidos nos mesmos dados experimentais é conhecida como **erro de conjunto** ou **erro de experimento**. Um experimento com três condições é um delineamento relativamente simples, assim, o efeito de realizar vários testes não é severo. Se aumentássemos o número de condições experimentais de três para cinco (apenas acrescentando mais dois grupos), o número de testes *t* necessário aumentaria para 10.[1] O valor do erro de conjunto poderia ser calculado utilizando a equação geral (8.1), na qual n é o número de testes executados com os dados. Com 10 testes executados, o valor do erro de conjunto seria 0,40 ($1 - 0{,}95^{10} = 0{,}40$), o que significa que teríamos uma probabilidade de 40% de ter cometido pelo menos um erro do Tipo I. Por essa razão, utilizamos a ANOVA em vez de realizar vários testes *t*:

$$\text{Erro de conjunto} = 1 - (0{,}95)^{n} \quad (8.1)$$

[1] Essas comparações são: 1 *vs.* 2, 1 *vs.* 3, *vs.* 4, 1 *vs.* 5, 2 *vs.* 3, 2 *vs.* 4, 2 *vs.* 5, 3 *vs.* 4, 3 *vs.* 5 e 4 *vs.* 5. O número de testes necessário é determinado utilizando a seguinte equação:

$$\text{Número de comparações } C = \frac{k!}{2(k-2)!}$$

onde *k* é o número de condições experimentais. O símbolo ! é denominado fatorial e significa que você deve multiplicar o valor que vem antes do símbolo por todos os números inteiros entre ele (inclusive) e zero (exclusive) (assim, $5! = 5\times4\times3\times2\times1 = 120$). Dessa forma, com cinco condições encontraremos que:

$$C = \frac{5!}{2(5-2)!} = \frac{120}{2\times6} = 10$$

Quando executamos testes *t*, testamos a hipótese de que as duas populações de onde as amostras foram retiradas apresentam a mesma média. Da mesma maneira, a ANOVA nos informa se três ou mais médias populacionais são iguais, assim, ela testa a hipótese de que as médias de todas as condições são iguais.

Uma ANOVA produz uma estatística F ou razão F que é semelhante a estatística t pois ela compara a variância sistemática nos dados com a variância não-sistemática. Contudo a ANOVA é um teste abrangente (*omnibus*), o que significa que ela testa um efeito experimental de forma global: assim, existem coisas que a ANOVA não nos informa. Embora a ANOVA nos informe se a manipulação experimental teve sucesso, ela não fornece informações específicas sobre quais grupos foram afetados. Supondo que um experimento tenha sido realizado com três grupos diferentes, a razão F simplesmente nos informa que as médias das três populações de onde os grupos foram retirados são diferentes (isto é, $m_1 = m_2 = m_3$ não é verdadeiro). Contudo, as médias podem diferir de várias maneiras. A primeira possibilidade é que as três médias sejam significativamente diferentes ($m_1 \neq m_2 \neq m_3$). Uma segunda possibilidade é que as médias das populações 1 e 2 não sejam diferentes mas tenham uma diferença significativa em relação à terceira população ($m_1 = m_2 \neq m_3$). Outra possibilidade é que as populações 2 e 3 não difiram mas sejam significativamente diferente da média da população 1 ($m_1 \neq m_2 = m_3$). Finalmente os grupos 1 e 3 podem indicar médias semelhantes mas o grupo 2 indica diferença significativa das populações 1 e 3 ($m_1 = m_3 \neq m_2$). Assim, a razão F informa apenas que a manipulação experimental tem algum efeito, mas não nos informa especificamente qual é esse efeito.

8.2.2 ANOVA como regressão ②

Muitos cientistas sociais não se dão conta de que a ANOVA e a análise de regressão são conceitualmente o mesmo procedimento.

Material com direitos autorais

O motivo é principalmente histórico, pois dois ramos distintos da metodologia se desenvolveram nas ciências sociais: a pesquisa correlacional e a experimental. Pesquisadores interessados em experimentos controlados adotaram a ANOVA como técnica estatística básica, ao passo que os interessados em determinar relacionamentos reais adotaram a regressão múltipla. Como os cientistas são pessoas maduras, inteligentes e racionais, nenhum dos grupos tentou difamar o outro defendendo que a sua escolha metodológica foi melhor. Com essa divisão, a metodologia tornou-se uma divergência entre os métodos estatísticos adotados pelos dois campos opostos (Crombach, 1957, documenta essa divisão em um ótimo artigo). Essa divisão vingou por muitas décadas, tanto que ainda hoje os estudantes aprendem regressão e ANOVA em contextos bem diferentes, e muitos livros didáticos ensinam ANOVA de uma forma inteiramente distinta da regressão.

Embora muitas pessoas mais inteligentes do que eu tenham tentado restabelecer essa conexão entre ANOVA e análise de regressão (como o grande Jacob Cohen, 1968), estou empolgado em escrever o meu – breve e tolo – esclarecimento (e colocar a bola em jogo nas Seções 5.10 e 7.8). Existem várias boas razões para pensar que a ANOVA deve ser ensinada no contexto da regressão. Primeiro, ela fornece um contexto familiar: desperdicei várias árvores explicando a regressão, então por que não utilizar essa base de conhecimento para explicar o novo conceito (tornando-o mais fácil de compreender). Segundo, o método tradicional de ensinar a ANOVA (conhecido como método da razão das variâncias) é bom para delineamentos simples, mas problemático em situações mais complexas (como na análise de covariância). O modelo de regressão se estende de uma forma lógica para esses delineamentos mais complexos sem que seja necessário ficar atolado na matemática. Finalmente, o método das razões das variâncias é extremamente não-manejável em circunstâncias incomuns, como quando você tem amostras de tamanhos diferentes.[2] O método da regressão torna essas situações muito mais simples. Embora essas razões sejam boas o suficiente, a verdade é que o SPSS também abandonou o método da razão das variâncias e passou a utilizar apenas o modelo de regressão (conhecido como modelo linear generalizado ou MLG). O resultado final das duas abordagens para a ANOVA é o mesmo e muitos textos já detalham a abordagem da razão das variâncias (recomendo o Howell, 2002). Assim faz sentido utilizar a abordagem da regressão.

A ANOVA é uma forma de comparar a razão entre a variância sistemática e a não sistemática em um estudo experimental. A razão dessas variâncias é conhecida como estatística F ou razão F. Quem leu o Capítulo 5 deve lembrar a razão F (veja a Seção 5.2.3) porque ela foi utilizada para avaliar quão bem um modelo de regressão pode prever uma saída comparada ao erro dentro do próprio modelo. Se você não leu o Capítulo 5 (provavelmente não!), dê uma olhada antes de ir adiante (deve demorar apenas duas semanas para lê-lo). A razão F na ANOVA é exatamente a mesma da regressão, exceto que um modelo de regressão contém somente previsores categóricos (isto é, variáveis de agrupamento). Desse modo, assim como o teste *t* pode ser representado por uma equação de regressão linear (veja a Seção 7.8), a ANOVA pode ser representada por uma equação de regressão múltipla na qual o número de previsores é um a menos do que o número de categorias da variável independente.

Vejamos um exemplo. Havia muita controvérsia sobre o medicamento Viagra quando escrevi a primeira edição deste livro. Agora existe bem menos polêmica, mas ela foi substituída por milhares de *spams* sobre o assunto (pelos quais eu serei grato, sem dúvida, daqui a 20 anos); assim, irei continuar com esse exemplo. O Viagra é um estimulante sexual (para tratar a impotência) utilizado com a crença de que ele fará do sujeito um amante melhor (estranha-

[2] Dito isso, vale tentar obter amostras de mesmo tamanho nas diferentes situações, pois delineamentos desbalanceados provocam complicações estatísticas (veja a Seção 8.2.9).

Material com direitos autorais

mente, vários jornalistas estão usando o produto em nome do "jornalismo investigativo"... hummm!). Suponha que vamos testar essa crença em três grupos de participantes e administrar a um grupo um placebo (como uma pílula de açúcar), a outro grupo uma dose baixa de Viagra e ao terceiro grupo uma dose alta. A variável dependente é uma medida objetiva da libido (apenas relatarei que ela foi mensurada ao longo de uma semana – o resto deixo para a imaginação de cada um). Os dados estão no arquivo **Viagra.sav** (descrito em detalhes posteriormente nesse capítulo) e são apresentados na Tabela 8.1.

Se quisermos prever os níveis da libido a partir dos diferentes níveis de Viagra, podemos utilizar a famosa equação geral:

$$\text{Saída}_i = (\text{Modelo}_i) + \text{erro}_i$$

Se você quiser utilizar um modelo linear, vimos na Seção 7.8 que quando existem somente dois grupos, podemos substituir o "modelo" nessa equação com uma equação de regressão linear como uma *dummy* (variável auxiliar) para descrever dois grupos-experimentais. Essa variável auxiliar é uma variável categórica com dois códigos numéricos (0 para um grupo e 1 para o outro). Com três grupos, no entanto, podemos estender essa ideia e utilizar um modelo e regressão múltipla com duas variáveis auxiliares. De fato, como regra geral, podemos estender um modelo para qualquer número de grupos e o número de variáveis auxiliares necessárias será um a menos que o número de categorias da variável independente. No caso de dois grupos, uma categoria foi considerada como base (lembre que na Seção 7.8 escolhemos a **condição foto** como uma base) e essa categoria foi codificado como 0. Quando existem três categorias também precisamos de uma categoria base, e deve ser escolhida a condição com a qual se pretende comparar os demais grupos. Em geral, essa categoria será o grupo-controle. Nos experimentos sociais bem-conduzidos quase sempre há um grupo de participantes que agiu como grupo-base para as demais categorias. Esse grupo atua como uma referência ou categoria base, embora o grupo que é escolhido dependa das hipóteses que se pretende testar. Em delineamentos não-balanceados (os grupos não são do mesmo tamanho), é importante que a categoria base contenha muitos casos para assegurar que as estimativas dos coeficientes de regressão sejam confiáveis. No exemplo do Viagra, podemos usar o grupo-placebo como categoria base porque esse é o grupo-controle. Estamos interessados em comparar tanto o grupo com altas doses quanto o com doses baixas com o grupo que não tomou Viagra. Se o grupo-placebo é a categoria base, as duas variáveis auxiliares que precisamos criar representam as duas outras condições, assim, devemos ter uma variável auxiliar denominada Alta e outra Baixa. A equação resultante é apresentada em (8.2):

$$\text{Libido}_i = b_0 + b_2\text{Alta}_i + b_1\text{Baixa}_i + \varepsilon_i \quad (8.2)$$

Na equação (8.2), a libido de uma pessoa pode ser prevista a partir do conhecimento do código do seu grupo (isto é, o código das variáveis auxiliares Alta e Baixa) e pelo intercepto (b_0) do modelo. As variáveis auxiliares na equação (8.2) podem ser codificadas de várias formas, mas a mais simples é utilizar uma técnica semelhante a do teste t. A categoria base é sempre codificada como 0. Se um participante recebeu uma alta dose de Viagra, ele é codificado como 1 para a variável auxiliar e 0 para todas as outras variáveis. Se um participante recebeu uma dose baixa de Viagra, ele será codificado como o valor 1 para a variável auxiliar baixa e como 0 para as demais (esse é o mesmo esquema utilizado na Seção 5.10). Utilizando

Tabela 8.1 Dados em **Viagra.sav**

	Placebo	Dose baixa	Dose alta
	3	5	7
	2	2	4
	1	4	5
	1	2	3
	4	3	6
$\bar{X}$	2,20	3,20	5,00
s	1,30	1,30	1,58
s^2	1,70	1,70	2,50
Média global = 3,467 Desvio padrão global = 1,767 Variância global = 3,124			

Material com direitos autorais

Tabela 8.2 Variáveis auxiliares para os três grupos do projeto experimental

Grupo	Variável auxiliar 1 alta y	Variável auxiliar 2 baixa y
Placebo	0	0
Dose baixa de Viagra	0	1
Dose alta de Viagra	1	0

esse esquema de codificação, podemos expressar cada grupo combinando os códigos das duas variáveis auxiliares (veja a Tabela 8.2).

Grupo-placebo: vamos examinar o modelo para o grupo-placebo. No grupo-placebo, as variáveis auxiliares dose alta e dose baixa são codificadas como 0. Além disso, se ignorarmos o termo erro (ε_i), a equação de regressão se torna:

$$\begin{aligned}\text{Libido}_i &= b_0 + (b_2 \times 0) + (b_1 \times 0)\\ \text{Libido}_i &= b_0\\ \overline{X}_{\text{Placebo}} &= b_0\end{aligned}$$

Essa é uma situação em que os grupos de dose alta e baixa foram ambos excluídos (eles foram codificados como 0). Queremos prever o nível de libido quando ambas as doses de Viagra forem ignoradas, desse modo, o valor previsto será a média do grupo-placebo (porque esse é o único grupo incluído no modelo). Assim, o intercepto do modelo de regressão, b_0, é sempre a média da categoria base (nesse caso, a média do grupo-placebo).

Grupo de alta dosagem: se examinarmos o grupo de alta dosagem, a variável auxiliar alta será codificada como 1 e a variável auxiliar baixa, como 0. Se substituirmos os valores desses códigos na equação (8.2), o modelo fica:

$$\begin{aligned}\text{Libido}_i &= b_0 + (b_2 \times 1) + (b_1 \times 0)\\ \text{Libido}_i &= b_0 + b_2\end{aligned}$$

Já sabemos que b_0 é a média do grupo-placebo. Se estivermos somente interessados no grupo de alta dosagem, então o modelo deve prever que o valor da libido para um dado participante iguale a média do grupo de alta dosagem. Dada essa informação, a equação torna-se:

$$\begin{aligned}\text{Libido}_i &= b_0 + b_2\\ \overline{X}_{\text{Alta}} &= \overline{X}_{\text{Placebo}} + b_2\\ b_2 &= \overline{X}_{\text{Alta}} - \overline{X}_{\text{Placebo}}\end{aligned}$$

Assim, b_2 representa a diferença entre as médias dos grupos de alta dosagem e o grupo-placebo.

Grupo de baixa dosagem: finalmente, se olharmos para o modelo quando uma dose baixa de Viagra for administrada, a variável auxiliar será codificada como 1 (e o grupo de alta dosagem como 0). Assim, a equação de regressão se torna:

$$\begin{aligned}\text{Libido}_i &= b_0 + (b_2 \times 0) + (b_1 \times 1)\\ \text{Libido}_i &= b_0 + b_1\end{aligned}$$

Sabemos que o intercepto é igual a média da categoria-base e que, para o grupo de baixa dosagem, o valor previsto deve ser a média da libido para o de baixa dosagem. Dessa forma, o modelo pode ser reduzido para:

$$\begin{aligned}\text{Libido}_i &= b_0 + b_1\\ \overline{X}_{\text{Alta}} &= \overline{X}_{\text{Placebo}} + b_1\\ b_1 &= \overline{X}_{\text{Baixa}} - \overline{X}_{\text{Placebo}}\end{aligned}$$

Assim, b_1 representa a diferença entre as médias dos grupos de baixa dosagem e o grupo-placebo. Essa forma de codificação com variáveis auxiliares é a mais simples, mas, como veremos mais tarde, existem outras formas em que as variáveis podem ser codificadas para testar hipóteses específicas. Esses esquemas de codificação alternativos são denominados *contrastes* (veja a Seção 8.2.10.2). O objetivo dos contrastes é codificar as variáveis auxiliares de tal forma que os valores-b representem diferenças entre os grupos que estamos interessados em testar.

Para ilustrar exatamente o que está acontecendo, criei um arquivo denominado **dummy.sav** na pasta do Chapter 8 disponível no *site* www.artmed.com.br. Esse arquivo contém os dados do Viagra, mas com duas variáveis auxiliares adicionais (**dummy1** e **dummy2**) que especificam a que grupo cada dado pertence (como na Tabela 8.2). Acesse esse arquivo

Material com direitos autorais

- Os coeficientes de regressão determinam a inclinação da linha; dessa forma, quanto maior os coeficientes, maior o desvio entre a linha e a média geral.
- Na ANOVA, os coeficientes de regressão são definidos pelas diferenças entre médias de grupos.
- Quanto maior as diferenças entre médias dos grupos, maior a diferença entre a linha de regressão e a média geral.
- Se as diferenças entre as médias dos grupos são grandes o suficiente, a linha resultante será um modelo melhor do que a média geral.
- Além disso, se as médias dos grupos são significativamente diferentes, o modelo de regressão deve se ajustar melhor do que a média geral.

Podemos ver se a última afirmação é verdadeira comparando a melhoria no ajuste devido ao uso do modelo (em vez de utilizar a média geral) com o erro que permanece. Outra forma de dizer isso é que quando a média geral é utilizada como um modelo, existirá certa quantidade de variação entre os dados e a média geral. Quando o modelo é ajustado ele irá explicar parte dessa variação, mas parte ainda permanecerá não-explicada. A estatística F é a razão entre a variação explicada e a não-explicada. Veja novamente a Seção 5.2.3 para refrescar a memória sobre esses conceitos antes de prosseguir. Isso pode parecer complicado, mas, de fato, quase tudo se resume a variações de uma equação simples (veja o Quadro 8.1).

8.2.4 Soma dos quadrados Total (SS_T) ②

Para encontrar a variação total dentro dos nossos dados calculamos as diferenças entre cada valor observado e a média geral (média de todos os valores). Elevamos então essas diferenças ao quadrado e adicionamos todas para obter a soma total dos quadrados (SS_T):

$$SS_T = \sum (x_i - \bar{x}_{geral})^2 \qquad (8.4)$$

Vimos também na Seção 1.4.1 que a variância e a soma dos quadrados estão relacionadas de forma que a variância $s^2 = SS_T/(N-1)$, onde N é o número de observações. Dessa forma, podemos calcular a soma dos quadrados total da variância de todas as observações (a **variância geral**) pela modificação da relação ($SS_T = s^2(N-1)$). A variância ge-

Quadro 8.1

Você pode achar surpreendente que a ANOVA se resume a uma equação (bem, mais ou menos) ②

A cada estágio da ANOVA, estamos avaliando variações (ou desvios) de um determinado modelo (seja ele o modelo mais básico ou o mais sofisticado). Vimos na Seção 1.4.1 que "o quanto" um modelo se desvia dos dados observados pode ser expresso, em geral, na forma da equação (8.3). Assim, na ANOVA, como na regressão, utilizamos a equação (8.3) para calcular a aderência ao modelo mais básico e depois a aderência ao melhor modelo (a linha da melhor aderência). Se o melhor modelo é mais adequado, ele deve se ajustar aos dados significativamente melhor do que o modelo básico:

$$\text{desvios} = \sum (\text{observado} - \text{modelo})^2 \qquad (8.3)$$

O ponto interessante é que todas as somas dos quadrados na ANOVA são variações dessa equação básica. O que muda é o que utilizamos como modelo e os dados que são ajustados. Olhe as várias seções das somas dos quadrados e compare as equações resultantes com a equação (8.3); espero que você consiga ver que todas são variações dessa equação básica!

Material com direitos autorais

ral é a variação entre todos os escores, independentemente das condições experimentais a que esses escores foram submetidos. Além disso, na Figura 8.1 ela será a soma das distâncias ao quadrado entre cada valor e a linha cheia horizontal. A variância geral para os dados do Viagra é apresentada na Tabela 8.1 e se contarmos o número de observações encontraremos que são 15 no total. Portanto, SS_T é determinado da seguinte forma:

$$\begin{aligned} SS_T &= s^2_{geral}(n-1) \\ &= 3{,}124(15-1) \\ &= 3{,}124 \times 14 \\ &= 43{,}74 \end{aligned}$$

Antes de prosseguirmos, é importante entender o que são graus de liberdade; assim, leia o Quadro 8.2.

8.2.5 Soma dos Quadrados do Modelo (SS_M) ②

Por enquanto, sabemos que a quantidade de variação entre os dados é 43,74 unidades. Agora precisamos saber quanto o modelo de regressão pode explicar dessa variação. Na ANOVA, o modelo é baseado nas diferenças entre as médias dos grupos, desse modo, a soma dos quadrados do modelo nos informa quanto da variação total pode ser explicado pelo fato de que dados diferentes provêm de grupos diferentes. Resumindo, o modelo representa o efeito da manipulação experimental.

Na Seção 5.2.3, vimos que a soma dos quadrados do modelo é calculada tomando-se as diferenças entre os valores previstos pelo modelo e a média geral (veja a Figura 5.3). Na ANOVA, os valores previstos pelo modelo são as médias dos grupos (portanto, na Figura 8.1 as linhas pontilhadas representam os valores da libido previstos pelo modelo). Para cada participante o valor previsto pelo modelo é a média do grupo a qual pertence o participante. No exemplo do Viagra, o valor previsto para os cinco participantes no grupo-placebo é 2,2, para os cinco participantes do grupo de baixa dosagem esse valor é 3,2 e para os cinco participantes do grupo de alta dosagem, 5,0. A soma dos quadrados do modelo requer que calculemos as diferenças entre o valor previsto de cada participante e a média geral. Essas diferenças são então elevadas ao quadrado e adicionadas (por razões que já devem estar claras

Quadro 8.2

Graus de liberdade ②

O conceito de graus de liberdade (gl) é difícil de explicar. Já o mencionei várias vezes ao longo do livro sem, no entanto, defini-lo. Muitos livros didáticos não explicam bem a ideia (embora Howell, 2002, p. 56, faça um bom trabalho) e este livro não será uma exceção à regra! Vou começar com uma analogia. Imagine que você seja o treinador de um time de rúgbi e tenha uma planilha do time com 15 células vazias que correspondem às posições dos jogadores no campo. Existe uma formação padrão nesse esporte: cada time tem 15 posições específicas que devem permanecer constantes durante o jogo. Quando o primeiro jogador chega existem 15 posições livres em que ele pode ser colocado. Você coloca o nome desse jogador em uma célula da planilha e o aloca em uma posição (por exemplo, linha média); uma posição no campo agora está ocupada. Quando o jogador seguinte chega, existem 14 posições livres e você ainda tem a liberdade de escolher em que posição o jogador será colocado. À medida que mais jogadores são alocados, chegará um ponto em que 14 posições já estarão preenchidas e restará um jogador. Não há liberdade de escolher a posição desse jogador, porque só existe uma possível. Portanto, existem 14 graus de liberdade, isto é, você tem possibilidade de escolher onde 14 jogadores irão jogar, mas para um jogador não há escolha. O grau de liberdade é um a menos que o número de jogadores.

(*Continua*)

Material com direitos autorais

Quadro 8.2 *(Continuação)*

Em termos estatísticos, os graus de liberdade estão relacionados ao número de observações que são livres para variar. Se retirarmos uma amostra de quatro observações de uma população para estimarmos o seu desvio padrão, precisaremos utilizar a média da amostra como uma estimativa da média da população. Assim, manteremos um valor constante. Digamos que a média da amostra seja 10; vamos assumir que a média da população seja esse valor e vamos mantê-lo constante. Com esse valor fixado, os quatro valores da amostra podem variar? A resposta é não porque para manter a média constante somente três valores podem variar. Por exemplo, se os valores da amostra forem 8, 9, 11 e 12 (média = 10) e trocarmos três desses valores por 7, 15 e 8, o último valor deve ser 10 para que a média não seja alterada. Portanto, se mantivermos um parâmetro constante, os graus de liberdade devem ser o tamanho da amostra menos um. Esse fato explica por que, quando utilizamos uma amostra para estimar o desvio padrão da população (como fizemos na Seção 1.4.1), devemos dividir a soma dos quadrados por $N - 1$ em vez de por N apenas. Para o SS_T, utilizamos a totalidade da amostra para determinar a soma dos quadrados, assim o total de graus de liberdade (GL_T) será o tamanho da amostra menos um (N – 1). Para os dados do Viagra, esse valor é 14.

para você a essa altura). Sabemos que o valor previsto para os participantes em um grupo específico é a média desse grupo. Portanto, a forma mais simples de calcular a SS_M é:

1. Calcular as diferenças entre a média de cada grupo e a média geral.
2. Elevar cada uma dessas diferenças ao quadrado.
3. Multiplicar cada resultado pelo número de participantes dentro de cada grupo (n_k).
4. Somar os valores de cada um dos grupos.

A expressão resultante desse processo é mostrada na equação (8.5):

$$SS_M = \sum n_k (\bar{x}_k - \bar{x}_{geral})^2 \qquad (8.5)$$

Utilizando as médias dos dados do Viagra, podemos calcular a SS_M da seguinte forma:

$$\begin{aligned} SS_M &= 5(2{,}200 - 3{,}467)^2 + 5(3{,}200 - 3{,}467)^2 \\ &\quad + 5(5{,}000 - 3{,}467)^2 \\ &= 5(-1{,}267)^2 + 5(-0{,}267)^2 + 5(1{,}533)^2 \\ &= 8{,}025 + 0{,}355 + 11{,}755 \\ &= 20{,}135 \end{aligned}$$

Para a SS_M, os graus de liberdade GL_M serão sempre menos um o número de parâmetros estimados menos um. Em resumo, o valor será o número de grupos menos 1 (que representaremos por $k - 1$). Assim, no caso de três grupos, o número de graus de liberdade será 2 (em virtude dos cálculos da soma dos quadrados ter por base a média dos grupos, duas serão livres para variar na população se a terceira for mantida constante).

8.2.6 Soma dos Quadrados dos Resíduos (SS_R) ②

Sabemos que existem 43,74 unidades de variação para serem explicadas nos nossos dados e que nosso modelo pode explicar 20,14 dessas unidades (aproximadamente a metade). A soma final dos quadrados é a soma dos quadrados dos resíduos (SS_R), que informa quanto da variação total não pode ser explicado pelo modelo. Esse valor é a quantidade de variação causada por fatores estranhos, como diferenças individuais de peso, testosterona ou outros. Se já conhecemos SS_T e SS_M, então a maneira mais fácil de calcular o valor de SS_R é subtraindo SS_M de SS_T ($SS_R = SS_T - SS_M$); contudo, fazendo isso teremos pouca ideia do que estamos calculando e, é claro, se errarmos no cálculo da SS_M ou da SS_T (ou de ambos), a SS_R também estará errada. Vimos na seção 5.2.3 que a soma dos quadrados dos resíduos é a diferença entre o que o modelo prevê e o

Material com direitos autorais

que é observado de fato. Sabemos que para um dado participante, o modelo prevê a média do grupo que esse participante pertence. Portanto, SS_R é calculado pela diferença entre o escore obtido por essa pessoa e a média do grupo que ela pertence. Em termos gráficos, as linhas verticais na Figura 8.1 representam essa soma dos quadrados. Essas distâncias entre cada dado e a média do grupo são elevadas ao quadrado e então somadas para fornecer a soma dos quadrados dos resíduos, SS_R (veja a equação (8.6)):

$$SS_R = \sum(x_{ik} - \bar{x}_k)^2 \qquad (8.6)$$

Agora, a soma dos quadrados para cada grupo representa a soma das diferenças ao quadrado entre o escore de cada participante no grupo e a média do grupo. Portanto, podemos expressar a SS_R como $SS_R = SS_{Grupo1} + SS_{Grupo2} + \ldots + SS_{Grupok}$. Dado que sabemos a relação entre a variância e a soma dos quadrados, podemos utilizar a variância de cada grupo dos dados do Viagra para criar uma equação como a que foi feita para a soma total dos quadrados. Dessa forma, a SS_R pode ser expressa pela equação (8.7):

$$SS_R = \sum s_k^2(n_k - 1) \qquad (8.7)$$

Isso significa que devemos pegar a variância de cada grupo (s_k^2) e multiplicá-la pelo número de pessoas naquele grupo menos um ($n_k - 1$). Depois de fazer isso para cada grupo, some todos os resultados. Para os dados do Viagra, teremos:

$$\begin{aligned} SS_R &= s^2_{\text{grupo }1}(n_1 - 1) + s^2_{\text{grupo }2}(n_2 - 1) \\ &\quad + s^2_{\text{grupo }3}(n_3 - 1) \\ &= (1{,}70)(5 - 1) + (1{,}70)(5 - 1) + (2{,}50) \\ &\quad (5 - 1) \\ &= (1{,}70 \times 4) + (1{,}70 \times 4) + (2{,}50 \times 4) \\ &= 6{,}8 + 6{,}8 + 10 \\ &= 23{,}60 \end{aligned}$$

O número de graus de liberdade para a quantidade SS_R (gl_R) pode ser obtido subtraindo-se o número de graus de liberdade do modelo do número de graus de liberdade total, isto é, $gl_R = gl_T - gl_M = 14 - 2 = 12$. Colocando de outra forma, isso é $N - k$, ou seja, o total da amostra (N) menos o número de grupos (k).

8.2.7 Médias ao quadrado ②

O valor SS_M nos informa quanto da variação o modelo de regressão (por exemplo, a manipulação experimental) explica e o SS_R informa quanto da variação pode ser creditado a fatores estranhos. Contudo, por que esses valores são somas, eles serão influenciados pelo número de valores que foram adicionados (por exemplo, SS_M utiliza a soma de apenas três valores diferentes – as médias dos grupos – comparadas a SS_R e SS_T, que utilizam a soma de 14 valores diferentes). Para eliminar esse viés, é possível calcular a soma dos quadrados média (conhecida como *média dos quadrados*, MS), que são simplesmente as somas dos quadrados divididas pelos graus de liberdade. Dividimos pelo número de graus de liberdade em vez de pelo número de parâmetros utilizados para calcular a SS porque estamos tentando extrapolar para uma população e alguns parâmetros dentro da população serão constantes. Assim, para os dados do Viagra, vamos obter as seguintes médias dos quadrados:

$$MS_M = \frac{SS_M}{gl_M} = \frac{20{,}135}{2} = 10{,}067$$

$$MS_R = \frac{SS_R}{gl_R} = \frac{23{,}60}{12} = 1{,}967$$

MS_M representa a quantidade média de variação explicada pelo modelo (isto é, a variação sistemática). Já MS_R é uma medida da quantidade média de variação explicada pelas variáveis estranhas (variação não-sistemática).

8.2.8 A razão F ②

A razão F é uma medida da variação explicada pelo modelo dividida pela variação explicada por fatores não-sistemáticos. Ela pode ser calculada pela divisão da média dos quadrados do modelo pela média dos quadrados dos resíduos:

Material com direitos autorais

$$F = \frac{MS_M}{MS_R} \quad (8.8)$$

Assim como o teste *t* independente, a razão F é uma medida da razão das variações sistemáticas para as não-sistemáticas. Como tal, ela é razão do efeito experimental para as diferenças individuais de desempenho. Um ponto interessante sobre a razão F é que, em virtude dela ser um quociente entre as variações sistemáticas e não-sistemáticas, se o seu valor for menor do que 1, então por definição ela representa um efeito não-significativo. Isso é verdade porque se a razão F é menor do que 1, significa que a MS_R é maior do que a MS_M, que em termos reais significa que existe mais variação não-sistemática do que sistemática. Você pode pensar nisso em termos do efeito de diferenças naturais na habilidade serem maiores do que as diferenças introduzidas pelo experimento. Nesse cenário, podemos estar seguros de que a nossa manipulação experimental não foi bem-sucedida (porque ela trouxe menos mudanças do que se tivéssemos deixado os participantes por si só!). Para os dados do Viagra a razão F é:

$$F = \frac{MS_M}{MS_R} = \frac{10,067}{1,967} = 5,12$$

Esse valor é maior do que 1, o que indica que a manipulação experimental teve um efeito além do efeito das diferenças de desempenho individuais. Contudo, ela não nos informa se a razão F é grande o suficiente para não ser um resultado casual. Para descobrir isso, devemos comparar o valor de F obtido contra o valor máximo que seria esperado apenas por acaso utilizando uma distribuição F com os mesmos graus de liberdade (esses valores podem ser encontrados no Apêndice A.3). Se o valor que obtivemos exceder esse valor crítico, podemos estar confiantes de que ele reflete um efeito da variável independente. Nesse caso, com 2 (gl_m) e 12 (gl_r) graus de liberdade, os valores críticos são 3,89 ($p = 0,05$) e 6,93 ($p = 0,01$), respectivamente. O valor observado, 5,12, é, dessa forma, significativo a 5%, mas não a 1%. Portanto, a significância exata fornecida pelo SPSS deve estar entre 0,05 e 0,01 (e de fato está).

8.2.9 Suposições da ANOVA ②

As suposições sob as quais a ANOVA é confiável são as mesmas dos testes paramétricos baseados na distribuição normal (veja a Seção 3.2). Isto é, os dados devem ser de uma população distribuída normalmente, as variâncias em cada uma das condições experimentais devem ser homogêneas, as observações devem ser independentes e a variável dependente deve ser mensurada pelo menos em uma escala de intervalo. Embora eu esteja sempre falando sobre a importância das hipóteses elas não são totalmente inflexíveis. Por exemplo, Lunney (1970) investigou o uso da ANOVA quando a variável dependente é dicotômica (podemos ter valores do tipo 0 e 1 apenas). Os resultados mostraram que quando os tamanhos dos grupos eram iguais, a ANOVA era precisa quando existiam pelo menos 20 graus de liberdade e a menor categoria continha no mínimo 20% de todas as respostas. Se a menor categoria de respostas continha menos do que 20% de todas as respostas, a ANOVA era precisa somente quando existiam 40 ou mais graus de liberdade. Esse estudo mostra que a ANOVA pode ser um procedimento bastante robusto.

Em termos de violação das hipóteses de homogeneidade da variância, a ANOVA é bastante robusta quando os tamanhos amostrais são iguais. Contudo, quando os tamanhos das amostras são diferentes, a ANOVA não é robusta a violações da homogeneidade da variância (por isso disse antes que vale a pena tentar coletar dados em grupos de mesmo tamanho!). Quando grupos com grandes tamanhos amostrais apresentam variâncias maiores do que grupos com tamanhos amostrais menores, a razão F resultante tende a ser conservadora. Quer dizer, é mais provável obter um resultado não-significativo mesmo quando existe uma diferença genuína na população. Já quando os grupos com grandes tamanhos amostrais apresentam variâncias menores do que os grupos com tamanhos amostrais menores, a razão F resultante tende a ser liberal. Ou seja, é mais provável obter um resultado significativo quando não existe uma diferença entre os grupos na população (ou seja, a taxa de Erro I não é controlada) – veja Glass, Peckham e Saunders (1972) para mais detalhes sobre o assunto.

Material com direitos autorais

isolado, ele não deve ser incorporado em qualquer outro contraste. Imagine a divisão da variância como fatiar um bolo. Você inicia com o bolo inteiro (a soma total dos quadrados) e depois corta o bolo em dois (SS_M e SS_R). Você então pega o pedaço que representa a SS_M e corta novamente em fatias menores. Uma vez que você tenha cortado uma fatia do bolo você não pode colocá-la de volta no bolo original, mas pode dividi-la em partes ainda menores. Da mesma forma, uma vez que a variância tenha sido dividida, ela não pode ser anexada a qualquer outra parte da variância; ela só pode ser novamente dividida em partes menores. Essa conversa sobre bolo está me deixando com fome, mas espero que tenha servido para ilustrar o que quero dizer.

Como uma segunda dica para selecionar contrastes, a regra da independência dos contrastes que acabei de explicar (a divisão do bolo!) significa que você deve sempre ter um contraste a menos que o número de grupos (assim, existirão $k - 1$ contrastes, onde k é o número de condições que estamos comparando). Finalmente, cada contraste deve comparar somente duas partes de variâncias. Essa regra final é a que permite que tiremos conclusões firmes sobre o que os contrastes nos dizem. A razão F nos informa que algumas das nossas médias diferem, mas não quais, e se executássemos um contraste em mais de duas partes de variâncias, enfrentaríamos o mesmo problema. Comparando apenas duas partes de variâncias, podemos estar seguros de que um resultado significativo representa uma diferença entre essas duas porções de variação experimental.

Em muitas pesquisas nas ciências sociais utilizamos pelo menos uma condição de controle e na grande maioria dos delineamentos experimentais previmos que as condições experimentais irão diferir da condição de controle (ou condições). Assim, **uma boa dica quando formos planejar comparações é comparar todos os grupos de condições experimentais com o grupo ou grupos-controle como sua primeira comparação.** Uma vez que tenhamos feito a primeira comparação, qualquer outra remanescente irá depender de quais grupos-experimentais você prevê que irão diferir.

Para ilustrar esses princípios, a Figura 8.4 e a Figura 8.5 mostram os contrastes que podem ser feitos com um experimento de quatro grupos. Observe que nos dois cenários existem três possíveis comparações (uma a menos do que o número de grupos). Também, cada contraste compara somente duas porções da variância. E mais, nos dois cenários o primeiro contraste é o mesmo: os grupos-experimentais são comparados contra o grupo-controle (ou grupos). Na Figura 8.4 existe somente uma condição de controle e assim essa porção da variância é utilizada somente no primeiro contraste (porque ela não pode ser mais dividida). Na Figura 8.5 existem dois grupos-controle e assim a porção da variância devido às condições de controle (contraste 1) pode ser dividida novamente como forma de verificar se os escores nos dois grupos-controle diferem entre si (contraste 3).

Na Figura 8.4, o primeiro contraste contém uma parte da variância que é acarretada pelos três grupos-experimentais e essa porção da variância é dividida verificando primeiramente se os grupos E1 e E2 diferem do E3 (contraste 2). É igualmente válido utilizar o contraste 2 para comparar os grupos E1 e E3 com E2, ou comparar os grupos E2 e E3 com E1. As comparações exatas que você escolher dependerão das suas hipóteses. Para o contraste 2 na Figura 8.4 ser válido precisamos ter uma boa razão para esperar que o grupo E3 seja diferente dos outros dois grupos. A terceira comparação na Figura 8.4 depende da que foi escolhida para o contraste 2. O contraste 2 necessariamente deve envolver a comparação de dois grupos-experimentais contra um terceiro e os grupos-experimentais escolhidos para serem combinados devem ser separados na comparação final. Finalmente, observe nas Figuras 8.4 e 8.5 que quando um grupo foi isolado em uma comparação ele não é mais utilizado em qualquer contraste posterior.

Quando executamos um contraste planejado, comparamos "porções" de variância e essas porções às vezes consistem em vários grupos. Às vezes é confuso entender exatamente o que esses contrastes nos informam. Quando

Material com direitos autorais

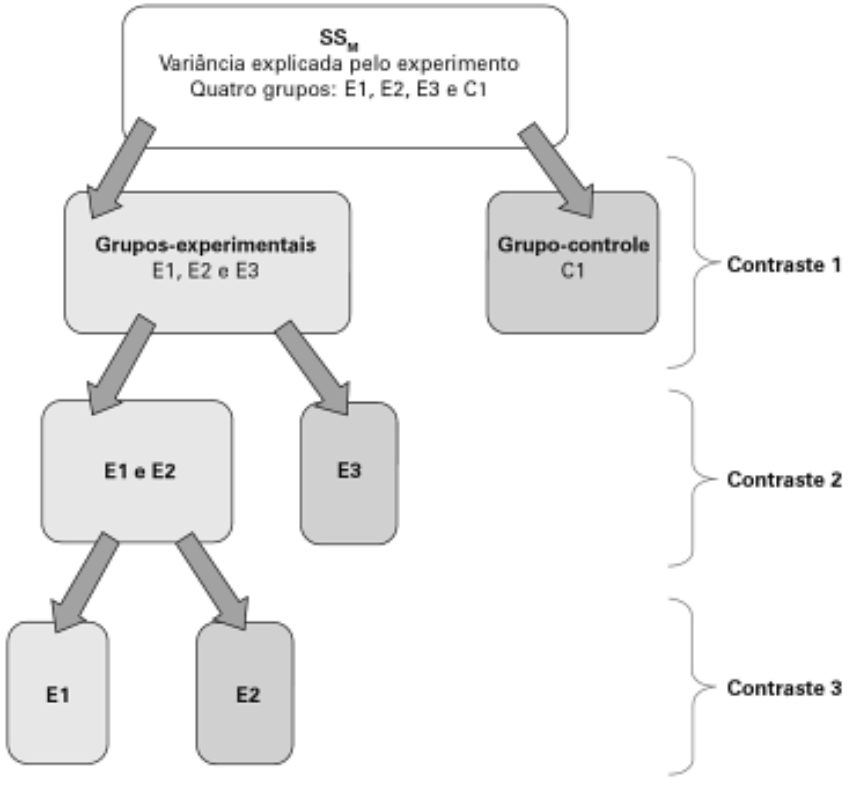


Figura 8.4 Partição da variância para comparações planejadas em um experimento de quatro grupos onde um é o grupo-controle.

você planeja um contraste que compara vários grupos a outro grupo, você está comparando as médias dos grupos em uma porção com a média do grupo em outra porção. Como exemplo, para os dados do Viagra eu sugeri que um primeiro contraste apropriado seria a comparação dos grupos das duas dosagens (alta e baixa) com o grupo-placebo. As médias dos grupos são 2,20 (placebo), 3,20 (de baixa dosagem) e 5,00 (alta dosagem); assim, a primeira comparação – que compara os dois grupos-experimentais contra o placebo – é comparar 2,20 (média do grupo-placebo) com a média dos outros dois grupos ((3,20 + 5,00)/2 = 4,10). Se esse primeiro contraste for significativo, podemos concluir que 4,10 é significativamente maior do que 2,20. Em termos do experimento, isso nos informa que a média desses grupos-experimentais é significativamente diferente da média do grupo-controle. Logicamente, isso significa que, se os erros padrão são os mesmos, o grupo-experimental com a média mais alta (o grupo de alta dosagem) será significativamente diferente da média do grupo-placebo. Contudo, o grupo-experimental com a média mais baixa (grupo de baixa dosagem) pode não diferir necessariamente do grupo-placebo; precisamos utilizar uma última comparação para esclarecer essas condições experimentais. Para os dados do Viagra, a comparação final é verificar se as duas condições experimentais diferem (isto é, as médias dos grupos de alta e baixa dosagem são significativamente diferentes). Se essa comparação for significativa,

Material com direitos autorais

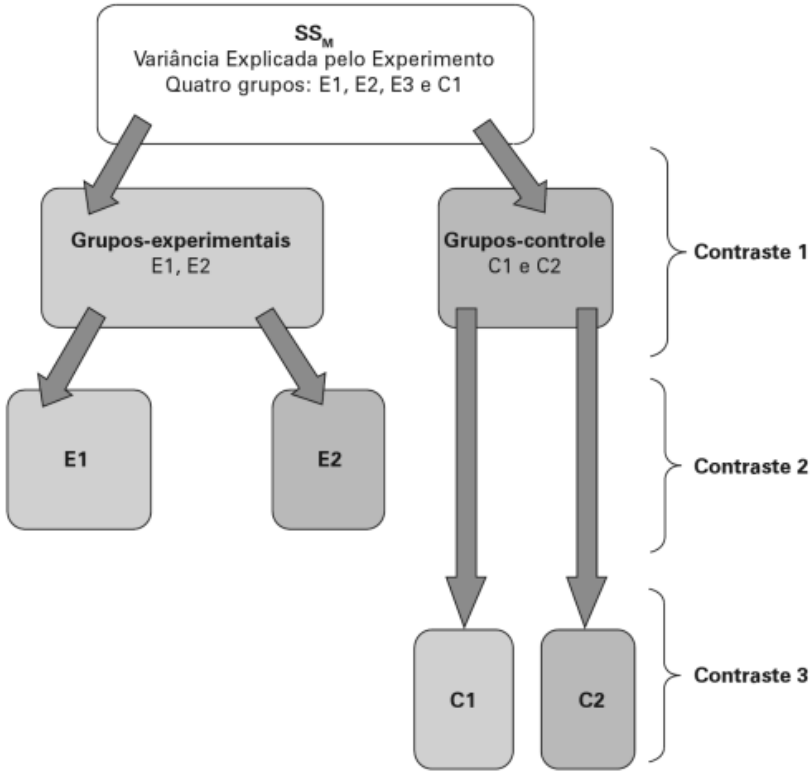


Figura 8.5 Partição da variância para comparações planejadas em um experimento de quatro grupos em que dois são grupos-controle.

podemos concluir que uma alta dose de Viagra afeta significativamente mais a libido do que uma baixa dosagem do medicamento. Se a comparação não for significativa, aceitamos que a dosagem do Viagra não afetou a libido. No primeiro caso, provavelmente as duas situações (alta e baixa dosagem) afetam a libido mais do que o placebo; já a segunda situação implica que uma dosagem baixa pode não ser diferente do que tomar o placebo. Contudo, a palavra *implica* é importante aqui: é possível que o grupo de baixa dosagem não difira do grupo-placebo. Para verificarmos isso, devemos executar um teste *post hoc*.

8.2.10.2 Definindo contrastes ponderados ③

Espero que a essa altura você já tenha alguma ideia de como planejar as suas comparações (se o seu cérebro ainda não explodiu!). Adoraria dizer a você que o difícil já passou e que o SPSS irá magicamente executar as comparações que você selecionou, mas isso não seria verdade. Para que o SPSS execute as comparações planejadas, você precisa informar quais grupos deverão ser comparados, e fazer isso pode ser complicado. De fato, quando executamos contrastes atribuímos valores a certas variáveis no modelo de regressão (desculpe, mas vou ter que falar sobre regressão de novo) – da mesma forma que fizemos quando utilizamos a codificação *dummy* (auxiliar) para a ANOVA principal. Para executar os contrastes, simplesmente atribuímos certos valores a variáveis auxiliares no modelo de regressão. Enquanto anteriormente definimos os grupos-experimentais atribuindo às variáveis auxiliares os valores 0 e 1, quando executamos

Material com direitos autorais

os contrastes utilizamos valores diferentes para especificar quais grupos queremos comparar. Os coeficientes resultantes no modelo de regressão (b_2 e b_1) representam as comparações nas quais estamos interessados. Os valores atribuídos as variáveis *dummy* (auxiliares) são conhecidos como pesos (ponderações).

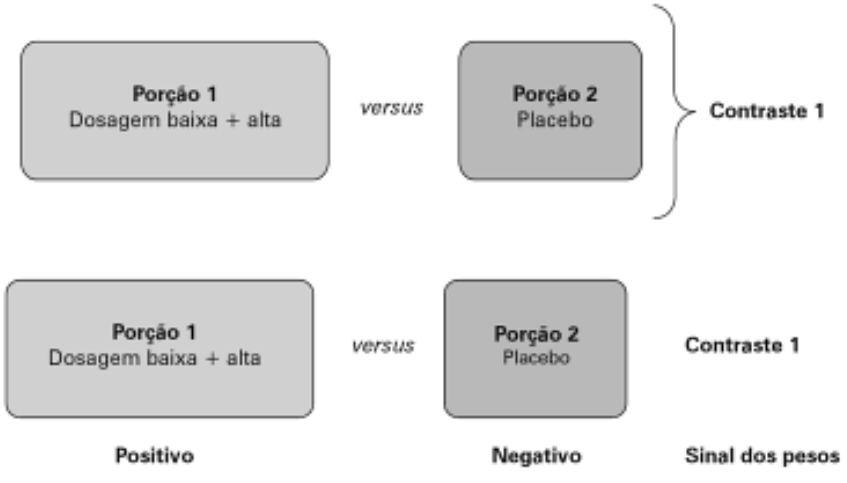
Esse procedimento é confuso, mas existem algumas regras básicas para atribuir valores às variáveis auxiliares a fim de obtermos as comparações desejadas. Explicarei essas regras antes de revelar como o processo funciona de fato. Lembre da seção anterior quando você leu essas regras e pense sobre o significado de "porção" de variação!

- **Regra 1**: Escolha comparações sensíveis. Lembre que você quer comparar somente duas porções de variação e que se um grupo é isolado em uma comparação, então esse grupo deve ser excluído de qualquer contraste subsequente.
- **Regra 2**: Grupos codificados com ponderações positivas serão comparados com grupos codificados com ponderações negativas. Assim, atribua a uma porção da variação ponderações positivas e à porção oposta, ponderações negativas.
- **Regra 3** A soma das ponderações (pesos) para uma comparação deve ser zero: se você somar os pesos para um determinado contraste o resultado deve ser zero.
- **Regra 4**: Se um grupo não está envolvido em uma comparação, atribua automaticamente o peso 0. Se for dado o peso 0 a um grupo, elimine esse grupo de qualquer outro cálculo.
- **Regra 5**: Para um dado contraste, os pesos atribuídos aos grupos em uma porção da variação deve ser igual ao número de grupos na porção contrária da variação.

Vamos usar algumas dessas regras a fim de determinar as ponderações para os dados do Viagra. A primeira comparação que escolhemos foi comparar os dois grupos-experimentais contra o grupo-controle:

Portanto, a primeira porção da variação contém os dois grupos-experimentais e a segunda parte contém somente o grupo-placebo. A regra 2 diz que devemos atribuir a uma porção ponderações positivas e a outra, negativas. Não importa a ordem em que isso é feito, mas por conveniência vamos atribuir à porção 1 uma ponderação positiva e a 2, negativa.

Utilizando a regra 5 os pesos atribuídos aos grupos na porção 1 devem ser equivalentes aos valores atribuídos aos grupos na porção 2. Existe somente um grupo na porção 2, assim, atribuímos a cada grupo na porção 1 o peso de 1. Da mesma forma, atribuímos ao grupo na porção 2 um peso igual ao número de grupos na porção 1. Há dois grupos na poção 1, portanto, atribuímos ao grupo-placebo peso 2.



Material com direitos autorais

Depois, combinamos os sinais dos pesos com a magnitude e chegamos a −2 (placebo), 1 (dosagem baixa) e 2 (dosagem alta):

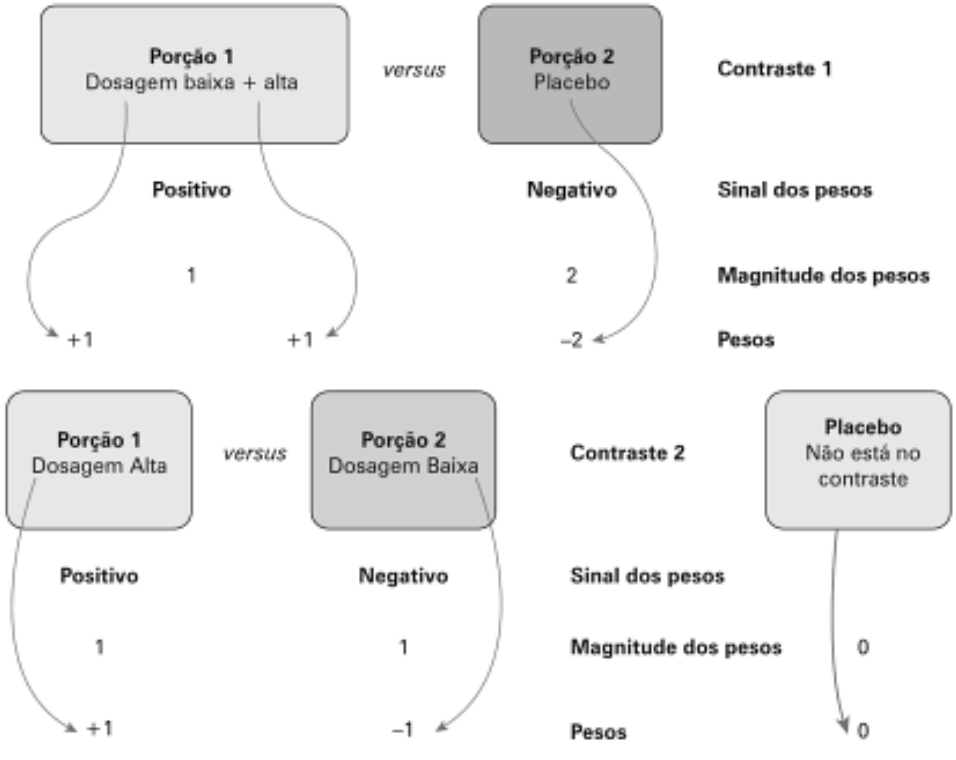
A regra 3 declara que para um dado contraste, a soma dos pesos deve ser zero; se as regras 2 e 5 forem seguidas, essa regra sempre será satisfeita (se você não seguiu adequadamente essas regras, isso irá aparecer quando os pesos forem somados). Vamos verificar somando os pesos: $1 + 1 - 2 = 0$.

O segundo contraste será comparar os dois grupos-experimentais e assim iremos ignorar o grupo-placebo. A regra 4 nos diz que devemos atribuir a esse grupo uma ponderação igual a 0 (porque dessa forma iremos eliminá-lo de qualquer cálculo). Restam ainda duas porções de variação: a porção 1, que contém a variação do grupo de baixa dosagem, e a porção 2, que contém a variação do grupo de alta dosagem. Seguindo as regras 2 e 5 devemos atribuir a um grupo o peso +1 e ao outro grupo o peso −1. O grupo-controle é ignorado (e assim ele terá um peso de 0). Se adicionarmos os pesos do contraste 2, veremos que a soma deverá ser novamente zero, isto é, $+1 - 1 + 0 = 0$.

Os pesos para cada contraste são codificados para duas variáveis *dummy* (auxiliares) na equação 8.2. Assim, esses códigos podem ser utilizados num modelo de regressão múltipla no qual b_2 representa o contraste 1 (comparando os grupos-experimentais ao grupo-controle), b_1 representa o contraste 2 (comparando os grupos de alta e baixa dosagem) e b_0 é a média geral (equação (8.9)):

$$\text{Libido}_i = b_0 + b_1\text{Contraste}_1 + b_2\text{Contraste}_2 \quad (8.9)$$

Cada grupo é especificado agora não pelo esquema de códigos 0 e 1 que foi inicialmente utilizado, mas pelo esquema de códigos dos dois contrastes. Um código de −2 para o primeiro contraste e um código de 0 para o segundo contraste identificam os participantes no grupo-placebo. Da mesma forma, o grupo de alta dosagem



Material com direitos autorais

res) **dummy1** e **dummy2** como variáveis previsoras (deixe todas as opções por padrão). A ANOVA principal para o modelo é a mesma de quando o código das variáveis auxiliares (*dummies*) foram utilizados; contudo, os coeficientes de regressão agora são outros.

A Saída 8.2 do SPSS mostra os resultados dessa regressão. Observe que o intercepto é a média geral, 3,467 (viu, eu não estou mentindo!). Segundo, o coeficiente da regressão para o contraste 1 é um terço da diferença entre a média das condições experimentais e a condição de controle (veja a seguir). Finalmente, o coeficiente da regressão para o segundo contraste é metade da diferença entre as médias dos grupos-experimentais (veja a seguir). Assim, quando uma comparação planejada é realizada na ANOVA, um teste *t* é conduzido comparando a média de uma porção da variação com a média de uma porção diferente. A partir das significâncias dos testes *t* podemos ver que nossos grupos-experimentais foram significativamente diferentes do grupo-controle ($p < 0,05$), mas que os grupos-experimentais não foram significativamente diferentes ($p > 0,05$).

8.2.10.3 Comparações não-ortogonais ②

Gastei bastante tempo explicando como projetar comparações ortogonais apropriadas sem mencionar as possibilidades oferecidas pelos contrastes não-ortogonais. Contrastes não-ortogonais são comparações que estão de alguma forma relacionadas e a melhor maneira de obtê-los é desobedecendo a regra número 1 da seção anterior. Utilizando a minha analogia do bolo, nas comparações não-ortogonais fatiamos o bolo e tentamos manter as fatias novamente juntas! Para o conjunto de dados do Viagra, por exemplo, um conjunto de contrastes não-ortogonais pode ter o mesmo contraste inicial (comparando os grupos-experimentais contra o grupo-controle), mas depois compara o grupo de baixa dosagem com o grupo-placebo. Isso viola a regra 1 porque o grupo-placebo está sozinho no primeiro contraste e foi utilizado novamente no segundo contraste. Os códigos para esse conjunto de contrastes são mostrados na Tabela 8.5; olhando para a última coluna fica claro que quando você multiplica e soma os códigos dos contrastes a soma não dá zero. Isso mostra que os contrastes não são ortogonais.

Nada existe de intrinsecamente errado em utilizarmos contrastes não-ortogonais. Contudo, se você escolher esse tipo de procedimento deve tomar muito cuidado em como irá interpretar os resultados. Com contrastes não-ortogonais as comparações feitas estão relacionadas, assim, a estatística teste resultante e o valor p estão correlacionados de alguma forma. Desse modo, você deve utilizar níveis de probabilidade mais conservadores

Saída 8.2 do SPSS

***Coefficients*[a] (Coeficientes)**

Model (Modelo)		*Unstandardized Coefficients* (Coeficientes não-padronizados)		*Standardized Coefficients* (Coeficientes padronizados)		
		B	*Std. Error* (Erro padrão)	*Beta*	*t*	*Sig.*
1	*(Constant)* (Constante)	3.467	0.362		9.574	0.000
	Dummy Variable 1 (Variável auxiliar 1)	0.633	0.256	0.525	2.474	0.029
	Dummy Variable 2 (Variável auxiliary 2)	0.900	0.443	0.430	2.029	0.065

a. Dependent Variable: Libido (Variável dependente: Libido)

Material com direitos autorais

Tabela 8.5 Contrastes não-ortogonais para os dados do Viagra

Grupo	Variável auxiliar 1 (Contraste$_1$)	Variável auxiliar 2 (Contraste$_2$)	Produto Contraste$_1$ × Contraste$_2$
Placebo	–2	–1	2
Baixa dosagem	1	0	0
Alta dosagem	1	1	1
Total	0	0	3

para aceitar que um dado contraste é significativo (veja a Seção 8.2.11). Por essa razão, recomendo que você utilize comparações planejadas ortogonais sempre que possível.

8.2.10.4 Contrastes padrão ②

Embora em geral você projete seus próprios contrastes, existem contrastes especiais para certas situações. Alguns desses contrastes são ortogonais enquanto outros não. Muitos procedimentos no SPSS permitem que você execute os contrastes mencionados nessa seção.

A Tabela 8.6 mostra os contrastes que estão disponíveis no SPSS para procedimentos tais como a regressão logística (veja a Seção 6.4.3), a ANOVA fatorial e a ANOVA de medidas repetidas (veja os Capítulos 9 e 10). Embora os códigos exatos não sejam fornecidos na Tabela 8.6, exemplos das comparações são feitos com as situações de três e de quatro grupos (onde os grupos são rotulados de 1, 2 e 3 e 1, 2,

Tabela 8.6 Contrastes padrão disponíveis no SPSS

Nome	Definição	Contraste	Três Grupos	Quatro Grupos
Desvios (Primeiro)	Compara o efeito de cada categoria (exceto a primeira) com o efeito experimental global	1	2 *vs.* (1, 2, 3)	2 *vs.* (1, 2, 3, 4)
		2	3 *vs.* (1, 2, 3)	3 *vs.* (1, 2, 3, 4)
		3		4 *vs.* (1, 2, 3, 4)
Desvios (último)	Compara o efeito de cada categoria (exceto a última) com o efeito experimental global	1	1 *vs.* (1, 2, 3)	1 *vs.* (1, 2, 3, 4)
		2	2 *vs.* (1, 2, 3)	2 *vs.* (1, 2, 3, 4)
		3		3 *vs.* (1, 2, 3, 4)
Simples (Primeiro)	Cada categoria é comparada com a primeira categoria	1	1 *vs.* 2	1 *vs.* 2
		2	1 *vs.* 3	1 *vs.* 3
		3		1 *vs.* 4
Simples (último)	Cada categoria é comparada com a última categoria	1	1 *vs.* 3	1 *vs.* 4
		2	2 *vs.* 3	2 *vs.* 4
		3		3 *vs.* 4
Repetido	Cada categoria (exceto a primeira) é comparada com a prévia	1	1 *vs.* 2	1 *vs.* 2
		2	2 *vs.* 3	2 *vs.* 3
		3		3 *vs.* 4
Helmert	Cada categoria (exceto a última) é comparada ao efeito médio de todas as categorias subsequentes	1	1 *vs.* (2, 3)	1 *vs.* (2, 3, 4)
		2	2 *vs.* 3	2 *vs.* (3, 4)
		3		3 *vs.* 4
Diferença (Helmert reversa)	Cada categoria (exceto a primeira) é comparada ao efeito médio de todas as anteriores	1	3 *vs.* (2, 1)	4 *vs.* (3, 2, 1)
		2	2 *vs.* 1	3 *vs.* (2, 1)
		3		2 *vs.* 1

Material com direitos autorais

3 e 4, respectivamente). Quando você codifica uma variável no editor de dados, o SPSS irá tratar o código de menor valor como o grupo 1, o próximo mais alto como grupo 2 e assim sucessivamente. Portanto, dependendo de quais comparações quer realizar, você deve codificar a variável de código de modo apropriado (e então utilizar a Tabela 8.6 como um guia sobre quais comparações o SPSS irá fazer). Os leitores espertos podem ter notado que alguns contrastes na Tabela 8.6 são ortogonais (os de Helmert e os contrastes diferenças) enquanto outros não são ortogonais (desvios, simples e repetidos). Você também pode ter notado que as comparações calculadas utilizando contrastes simples são as mesmas obtidas pela utilização das variáveis auxiliares na Tabela 8.2.

8.2.10.5 Contrastes polinomiais: análise da tendência ②

Um tipo de contraste deliberadamente omitido na Tabela 8.6 é o polinomial. Esse contraste testa tendências nos dados e na sua forma mais básica ele procura por tendências lineares (isto é, se a média dos grupos aumenta de forma proporcional). Contudo, existem tendências mais complexas – como a quadrática, cúbica e quártica – que também podem ser examinadas. A Figura 8.6 mostra diagramas de tipos de tendência que podem existir em um conjunto de dados. A **tendência linear** representa uma mudança proporcional no valor da variável dependente ao longo de categoriais ordenadas. Uma **tendência quadrática** é quando existe uma mudança na direção de uma linha (por exemplo, a linha é curva em um lugar). Um exemplo disso é quando um medicamento melhora o desempenho em uma tarefa no início, mas à medida que a dose aumenta o desempenho volta a cair. É óbvio que para achar uma tendência quadrática são necessários pelo menos três grupos (porque na situação de dois grupos não existem categorias suficientes da variável independente para que as médias da variável dependente mudem de uma forma e depois de outra). Uma **tendência cúbica** é quando existem duas mudanças na direção da tendência. Por exemplo, a média da variável dependente primeiro aumenta com as primeiras categorias da variável independente, depois ao longo de outras categorias as mé-

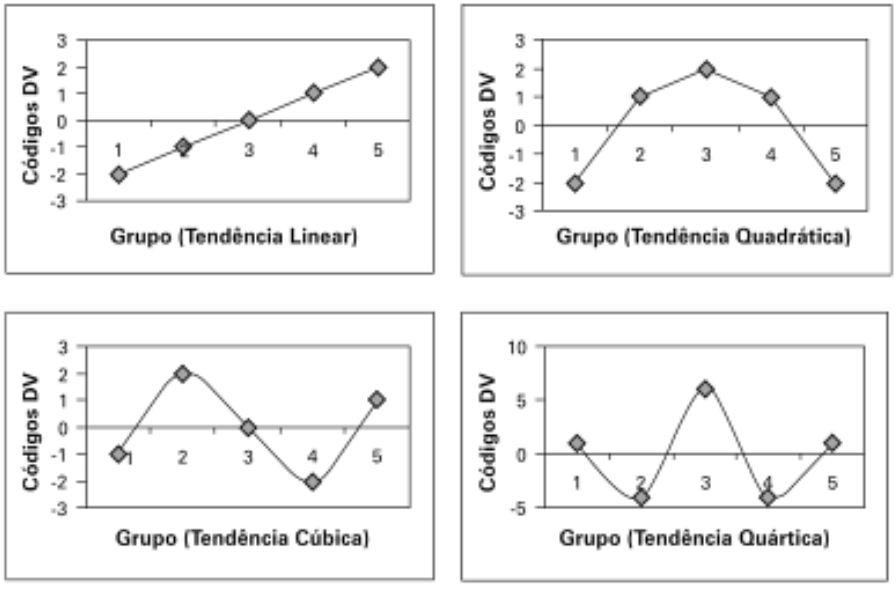


Figura 8.6 Tendências linear, quadrática, cúbica e quártica em cinco grupos.

dias diminuem, mas nas últimas categorias as médias se elevam novamente. Para ter duas mudanças na direção da média você deve ter pelo menos quatro categorias da variável independente. A última tendência com a qual você pode se deparar é a **quártica**, que apresenta três mudanças de direção (portanto, serão necessárias pelo menos cinco categorias da variável independente).

Tendências polinomiais devem ser examinadas em conjuntos de dados em que faz sentido ordenar as categorias da variável independente (por exemplo, se você administrou cinco doses de uma droga, faz sentido examinar as cinco doses na sua ordem de magnitude). Para os dados do Viagra existem somente três grupos, desse modo, podemos esperar encontrar somente uma tendência linear ou quadrática (e não teria sentido testar para tendências de ordens mais altas).

Cada uma dessas tendências apresenta um conjunto de códigos para as variáveis auxiliares no modelo de regressão, portanto, faremos o mesmo que foi feito para os contrastes planejados exceto que os códigos já foram escolhidos para representar o tipo de tendência de interesse. De fato, os diagramas na Figura 8.6 foram construídos traçando os valores dos códigos para os cinco grupos. Na verdade, se você adicionar códigos para uma dada tendência, a soma será igual a zero, e se os códigos forem multiplicados você vai encontrar que a soma dos produtos também será igual a zero. Dessa forma, esses contrastes são ortogonais. A vantagem desses contrastes é que não é necessário construir uma codificação própria, pois esses códigos já existem.

8.2.11 Procedimentos *post hoc* ②

Frequentemente, você não tem previsões específicas *a priori* sobre os dados coletados e, em vez disso, está interessado em explorar os dados para verificar a existência de quaisquer possíveis diferenças entre médias. Esse procedimento é, às vezes, denominado *mineração de dados* ou *análise de dados exploratória*. Eu acho que esses dois termos têm uma conotação de "manipulação" dos dados e prefiro pensar nesses procedimentos como "encontrar as diferenças que eu teria previsto se fosse esperto o suficiente".

Testes *post hoc* consistem em **comparações em pares** planejadas para comparar todas as diferentes combinações dos grupos sendo testados. Assim, é como tomar todos os pares de grupos possíveis e executar um teste *t* em cada um. Pode parecer idiota explicar isso (mas eu sou um pouco idiota) considerando o que eu já disse sobre o problema do crescimento do erro de conjunto. Contudo, as comparações aos pares controlam o erro de conjunto corrigindo o nível de significância de cada teste de forma que o erro do Tipo I total (α) ao longo de todas as comparações permaneça ao nível especificado (0,05). Existem várias formas de controlar o erro de conjunto. A mais popular (e mais fácil) é dividir o valor de α pelo número de comparações, assegurando, dessa forma, que o erro do Tipo I acumulado fique abaixo do valor especificado (0,05). Portanto, se forem executados 10 testes, utilizamos 0,005 como nosso critério de significância. Esse método é conhecido como **correção de Bonferroni** em homenagem a Carlo Bonferroni (Figura 8.7). Existe uma contrapartida no controle da taxa de erro de conjunto que é a perda de poder estatístico. Isso quer dizer que a probabilidade de rejeitar um efeito que existe aumenta (o erro do Tipo II). Se formos conservadores na taxa de erro do Tipo I para cada comparação, aumentamos a chance de deixar passar uma diferença genuína nos dados.

Contudo, quando consideramos procedimentos *post hoc* precisamos levar em conta três coisas: (1) o teste controla a taxa de erro do Tipo I; (2) o teste controla a taxa de erro do Tipo II (isto é, o teste apresenta um bom poder estatístico); e (3) o teste é confiável quando as suposições da ANOVA são violadas?

Eu adoraria discutir todos os detalhes tediosos dos tipos de testes *post hoc*, mas não há razão para tanto. Existem alguns textos excelentes disponíveis para quem quer saber mais (por exemplo, Toothaker, 1993; Klokars e Sax, 1986) e o SPSS também fornece 18 tipos de teste *post hoc*, portanto, eu teria que des-

Material com direitos autorais

Figura 8.7 Carlo Bonferroni, antes que a fama da sua correção o levasse à bebida, às drogas e ao fanatismo pela estatística.

matar muitos quilômetros de floresta para explicá-los. O importante é que você saiba quais testes *post hoc* executam melhor de acordo com os três critérios mencionados acima.

8.2.11.1 Procedimentos *post hoc* e erros do Tipo I (α) e Tipo II (β) ②

A taxa de erro do Tipo I e o poder estatístico de um teste estão conectados. Portanto, existe sempre uma contrapartida: se o teste é conservador (a probabilidade de erro do Tipo I é pequena), é provável que ele tenha um baixo poder estatístico (a probabilidade de erro do Tipo II será alta). Assim, é importante que os procedimentos de comparações múltiplas controlem a taxa de erro do Tipo I, mas sem uma perda substancial no poder. Se um teste é muito conservador, provavelmente iremos rejeitar diferenças entre médias que são, na realidade, significativas.

A comparação de pares MDS (Menor Diferença Significativa) (*LDS – Least-Significant Difference*) não tenta controlar o erro do Tipo I e é equivalente à execução de múltiplos testes t. A única diferença é que a MDS requer que a ANOVA global seja significativa. O procedimento ENK (Estudentizado de Newman-Keuls) também é um teste bastante liberal e não controla o erro de conjunto. Os testes de Bonferroni e Tukey controlam a taxa de erro do Tipo I muito bem, mas são testes conservadores (eles têm poder estatístico deficiente). O de Bonferroni apresenta um poder maior quando o número de comparações é pequeno, já o de Tukey é mais poderoso quando testa uma grande quantidade de médias. O teste de Tukey geralmente tem um poder maior do que os testes de Dunn e Scheffé. O procedimento de Ryan, Einot, Gabril e Welsch Q (REGWQ) apresenta um bom poder e um controle satisfatório do erro do Tipo I. De fato, quando queremos testar todos os pares de médias esse procedimento é provavelmente o melhor. Contudo, quando os tamanhos dos grupos são diferentes, esse procedimento não deverá ser utilizado.

8.2.11.2 Procedimentos *post hoc* e violações das hipóteses do teste ②

A maioria das pesquisas sobre testes *post hoc* tem se debruçado no problema do desempenho do teste quando os tamanhos dos grupos são diferentes (delineamento desbalanceado), quando as variâncias das populações são muito distintas e quando os dados não são normalmente distribuídos. A boa notícia é que muitos procedimentos de comparações múltiplas tem um bom desempenho sob pequenos desvios da normalidade. A má notícia é que elas apresentam um mau desempenho quando os tamanhos dos grupos são diferentes e quando as variâncias populacionais são desiguais.

O teste GT2 de Hochberg e de Gabriel foram projetados para lidar com situações em que os tamanhos amostrais são diferentes. O procedimento de Gabriel é geralmente mais poderoso, mas pode se tornar muito liberal quando os tamanhos das amostras são bem diferentes. O GT2 de Hochbert não é confiável quando as variâncias populacionais são diferentes e deve ser utilizado somente quando você estiver seguro de que esse não é o caso. Existem várias comparações

Material com direitos autorais

derações ele deve atribuir para cada grupo. O primeiro passo é decidir quais comparações queremos fazer e então quais pesos devem ser atribuídos a cada grupo para cada um dos contrastes. Já vimos esse processo na Seção 8.2.10.2, assim, sabemos que os pesos para o primeiro contraste são −2 para o grupo-placebo, 1 para o grupo de baixa dosagem e novamente 1 para o grupo de alta dosagem. Iremos especificar esse contraste primeiro. É importante estar seguro de que o peso correto seja atribuído para cada grupo, desse modo, você deve lembrar que o primeiro peso atribuído deve ser a ponderação do primeiro grupo (isto é, o grupo codificado com o menor valor no editor de dados). Para os dados do Viagra, o grupo codificado com o menor valor é o placebo (que tem o código 1) e devemos entrar com a ponderação para esse grupo primeiro. Clique no espaço denominado **Coefficients** (Coeficientes) com o mouse e digite "-2" e clique em Add (Adicionar). O passo seguinte é entrar com a ponderação para o segundo grupo, que para os dados do Viagra é o grupo de baixa dosagem (porque esse grupo está codificado no editor com o segundo valor mais alto). Clique novamente no espaço denominado **Coefficients** (Coeficientes) com o mouse e digite "1"clicando em seguida em Add (Adicionar). Finalmente, devemos entrar com a ponderação do último grupo que, para os dados do Viagra, é o grupo de alta dosagem (no editor esse grupo foi codificado com o valor mais alto). Clique em **Coefficients** (Coeficientes) com o mouse e digite "1" e, depois, clique em Add (Adicionar). O resultado deve ser semelhante ao da Figura 8.10.

Uma vez que os pesos foram atribuídos você poderá alterar ou remover qualquer um deles utilizando o mouse para selecionar a ponderação que quer mudar. O peso irá aparecer no espaço denominado **Coefficients** (Coeficientes), onde você poderá digitar o novo valor e então clicar em Change (Mudar). Alternativamente, você pode clicar em qualquer um dos pesos e removê-lo completamente clicando em Remove (Remover). Utilizando as ponderações, o SPSS calcula o coeficiente total

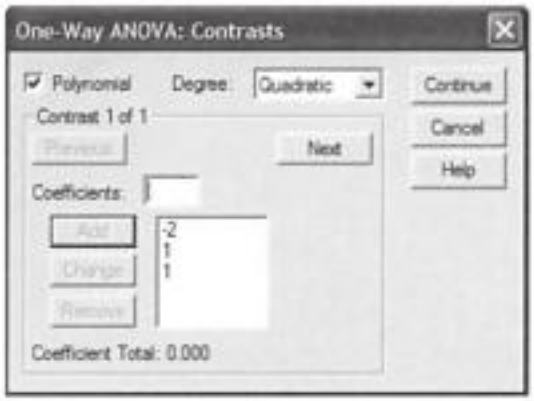


Figura 8.10 Caixa de diálogo dos contrastes completada para o primeiro contraste para os dados do Viagra.

que, como vimos na Seção 8.2.10.2, deve ser igual a zero. Se esse coeficiente for qualquer outro valor que não zero, você deve retornar e verificar se todas as regras para a atribuição de pesos foram seguidas.

Uma vez que você tenha especificado o primeiro contraste, clique em Next (próximo). Os pesos que você recém digitou desaparecerão e a caixa de diálogo estará pronta para a entrada do segundo contraste que, nesse caso, é o último contraste. Sabemos da Seção 8.2.10.2 que os pesos para o segundo contraste foram: 0 (para o grupo-placebo), +1 para o grupo de baixa dosagem e −1 para o grupo de alta dosagem. Podemos especificar esse contraste da mesma forma que antes. Lembrando que a primeira ponderação a ser entrada é a do grupo-placebo, devemos entrar o valor 0 como a primeira ponderação. Clique no espaço denominado **Coefficients** (Coeficientes) com o mouse e digite "0" e clique em Add (Adicionar). A seguir precisamos entrar com a ponderação para o grupo de baixa dosagem clicando novamente no espaço **Coefficients** (Coeficientes), digitando "1" e clicando em Add (Adicionar). Finalmente, será adicionada a ponderação para o grupo de alta dosagem da mesma forma só que agora com o cuidado de digitar o valor "-1". O resultado deve ser uma caixa de diálogo semelhante à da Figura 8.11. Note que a soma dos pesos é zero, como ocorreu com o contraste 1. Lembre-se de en-

Material com direitos autorais

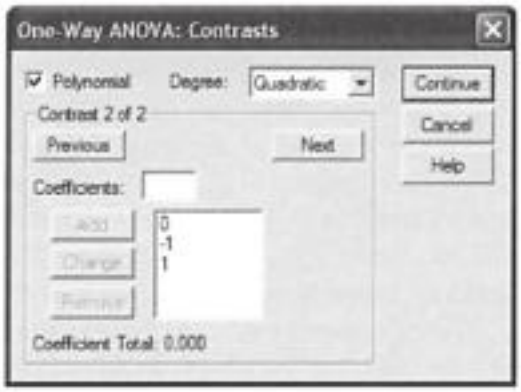


Figura 8.11 Caixa de diálogo dos contrastes preenchida para o segundo contraste para os dados do Viagra.

trar com o peso 0 para qualquer grupo que não estiver no contraste. Quando todos os contrastes planejados tiverem sido especificados, clique em (Continuar) para retornar à caixa de diálogo principal.

8.3.2 Testes *Post Hoc* no SPSS ②

Depois de informar ao SPSS quais as comparações planejadas a serem feitas, podemos escolher realizar testes *post hoc*. Teoricamente, se fizemos as comparações planejadas não precisamos realizar testes *post hoc* (porque as hipóteses de interesse já foram testadas). Da mesma forma, se tivéssemos escolhido realizar testes *post hoc* não seria necessário fazer os contrastes planejados (porque não temos hipóteses para testar). Contudo, para poupar espaço, iremos realizar alguns testes *post hoc* para os dados do Viagra. Clique em na caixa de diálogo principal para acessar a caixa de diálogo para os testes *post hoc* (Figura 8.12).

Na Seção 8.2.11.3, recomendei vários procedimentos *post hoc* para várias situações. Para os dados do Viagra, os tamanhos das amostras são iguais e não existe necessidade de realizar o teste de Gabriel. Devemos utilizar o teste de Tukey e o REGWQ e checar os resultados com o procedimento de Games-Howell. Temos uma hipótese específica que tanto o grupo de alta dosagem quanto o de baixa dosagem devem diferir do grupo-placebo e assim podemos utilizar o teste de Dunnett para examinar essas hipóteses. Uma vez que você tenha selecionado o teste de Dunnett, você pode mudar a categoria de controle (o padrão é utilizar a última categoria) para especificar que a primeira categoria será utilizada como categoria de controle (porque o grupo-placebo foi codificado com o valor mais baixo). Você pode, também, escolher se irá realizar um teste do tipo bilateral ( ) ou unilateral. Se você escolher o teste unilateral, deve prever se você acredita que a média do grupo-controle será menor do que a média de um grupo-expe-

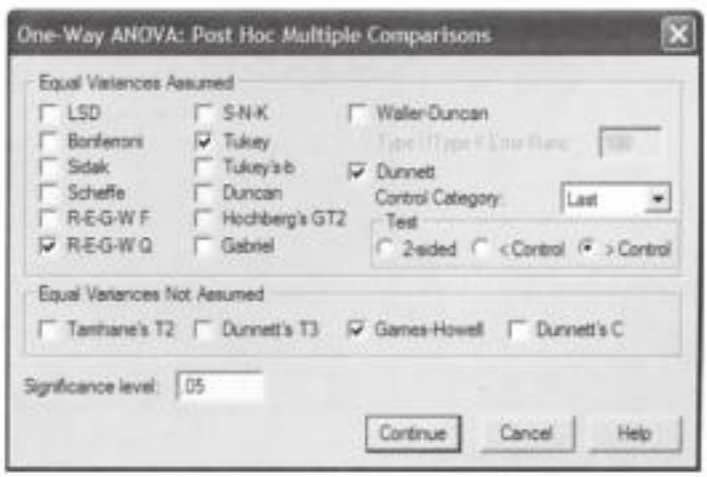


Figura 8.12 Caixa de diálogo para especificar testes *post hoc*.

Material com direitos autorais

rimental em particular (Control) ou maior do que a média de um grupo-experimental em particular (Control). Esses são todos os testes *post hoc* que você necessita especificar e quando a caixa de diálogo estiver completa ela deverá ser semelhante à da Figura 8.12. Clique em Continue (continuar) para voltar à caixa de diálogo principal.

8.3.3 Opções ②

As opções para a ANOVA de um fator são bem simples. Primeiro você pode solicitar estatísticas descritivas, o que irá produzir uma tabela de médias, desvios padrão, erros padrão, amplitudes e intervalos de confiança para as médias de cada grupo. Essa é uma opção útil, pois poderá auxiliar na interpretação dos resultados finais. Uma opção crucial a ser selecionada é o teste de homogeneidade das variâncias. Como ocorre com os testes t, há uma suposição de que as variâncias dos grupos são iguais e selecionando essa opção essa hipótese é testada. O SPSS utiliza o teste de Levene, que testa a hipótese de que as variâncias de cada grupo são iguais (veja a Seção 3.6). Se a hipótese de homogeneidade das variâncias for violada, o SPSS oferece duas versões alternativas da razão F: a de Brow-Forsythe (1974) e a de Welch (1951). Essas duas estatísticas são discutidas no Quadro 8.3 (e, se você está realmente entediado, olhe o arquivo **WelchF.pdf** no *site* www.artmed.com.br), mas é suficiente dizer que elas devem ser selecionadas apenas como garantia para o caso de a hipótese ser violada.

Existe, também, a opção de ter um diagrama das médias (selecionar a opção **Means plot**) (Figura 8.13). Se essa opção for selecionada, um diagrama de linha do grupo de médias será apresentado na saída. Finalmente, as opções nos deixam especificar se queremos excluir casos na base **listwise** ou em uma base **per analysis**. Essa opção é útil se você está realizando várias ANOVAS com diferentes variáveis dependentes. A primeira opção (**Exclude cases analysis by analysis**) (Excluir casos análise por análise) exclui qualquer caso que tiver um valor faltando (*missing*) tanto para a variável independente

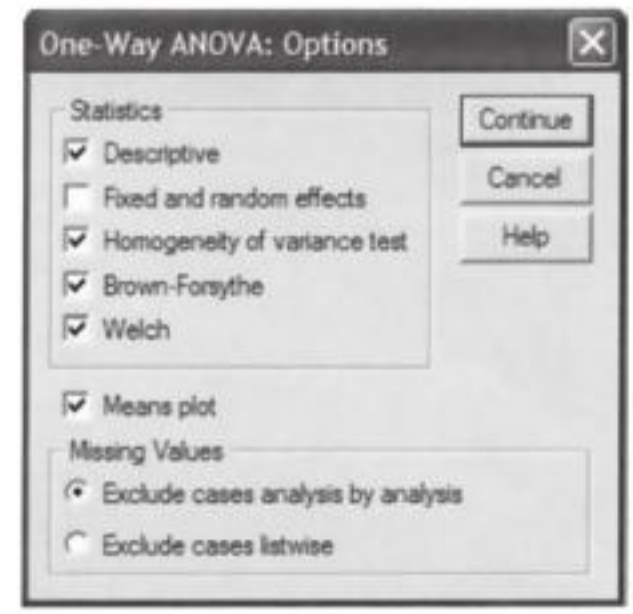


Figura 8.13 Opções para a ANOVA de um fator.

quanto para a variável dependente utilizadas nessa análise em particular. A opção **Exclude cases listwise** (Excluir casos por lista) excluirá de todas as análises qualquer caso que tenha um valor faltando para a variável independente ou qualquer variável dependente especificada. Assim, se você se acostumar a não realizar centenas de ANOVAs em diferentes variáveis (veja o Capítulo 14 sobre MANOVA) os valores padrão serão adequados. Quando você tiver selecionado as opções apropriadas, clique em Continue (continuar) para retornar à caixa de diálogo principal e clique em OK para rodar a análise.

8.4 SAÍDAS DA ANOVA DE UM FATOR (*ONE WAY*) ②

8.4.1 Saídas para a análise principal ②

Se você carregar os dados do Viagra (ou digitá-los) e selecionar todas as opções que eu sugeri, deverá verificar que a saída será semelhante a que segue. Se a sua saída for diferente não se assuste porque um de nós fez algo errado – espero que não eu. A Figura 8.14 mostra um diagrama de barras de erro dos dados do Viagra com uma linha superposta para apresentar a tendência geral das médias ao longo dos grupos. Esse gráfico não é automaticamente produzido pelo SPSS; contudo, um diagrama de linha será produzido se a opção **Means plot** (Representar médias) for selecionada (esse grá-

Quadro 8.3

O que eu faço na ANOVA quando a hipótese da homogeneidade das variâncias é violada? ③

Na Seção 8.2.9.1, mencionei que quando os tamanhos dos grupos são desiguais, violações na hipótese de homogeneidade das variâncias podem ter sérias consequências. O SPSS incorpora opções para duas razões F, que foram projetadas para serem mais robustas caso a homogeneidade das variâncias não seja válida. A primeira é a razão F de Brown e Forsythe (1974), que é bem fácil de explicar. Relatei anteriormente que quando os tamanhos dos grupos são diferentes e os grupos maiores apresentam as maiores variâncias, isso introduz um viés na razão F, que tende a ser mais conservadora. Se você voltar à equação (8.7), isso ficará claro, pois para calcular SS_R as variâncias são multiplicadas pelos seus tamanhos amostrais (menos 1), assim, nessa situação obteremos um grande tamanho amostral multiplicado por uma grande variância que irá inflar o valor do SS_R. Que efeito isso vai causar na razão F? Bem, a razão F é proporcional a SS_M/SS_R, desse modo, se SS_R é grande, então a razão F será pequena (o motivo para ser mais conservadora: seu valor está sendo reduzido demais!). Brown e Forsythe contornaram esse problema ponderando as variâncias dos grupos, não pelos seus tamanhos amostrais, mas pelo inverso dos seus tamanhos amostrais (na verdade, eles utilizam n/N, o tamanho da amostra como uma proporção do tamanho total das amostras). Isso significa que o impacto de grandes tamanhos amostrais com grandes variâncias será reduzido:

$$F_{BF} = \frac{SS_M}{SS_{R_{BF}}} = \frac{SS_M}{\sum s_k^2\left(1-\frac{n_k}{N}\right)}$$

Assim, para os dados do Viagra, a SS_M é a mesma de antes (20,135), dessa forma, a equação fica:

$$\begin{aligned} F_{BF} &= \frac{20{,}135}{s^2_{grupo1}\left(1-\frac{n_{grupo1}}{N}\right)+s^2_{grupo2}\left(1-\frac{n_{grupo2}}{N}\right)+s^2_{grupo3}\left(1-\frac{n_{grupo3}}{N}\right)} \\ &= \frac{20{,}135}{1{,}70\left(1-\frac{5}{15}\right)+1{,}70\left(1-\frac{5}{15}\right)+2{,}50\left(1-\frac{5}{15}\right)} \\ &= \frac{20{,}135}{3{,}933} \\ &= 5{,}119 \end{aligned}$$

Essa estatística é avaliada utilizando os graus de liberdade do modelo e dos termos erro. Para o modelo, o gl_M é o mesmo de antes (isto é, $k - 1 = 2$), mas é feito um ajuste nos graus de liberdade dos resíduos, gl_R. Os cálculos desse valor bem como do F de Welch (1951) estão além dos objetivos deste livro, mas se você estiver interessado consulte o apêndice A.5 (ou o arquivo **WelchF.pdf** no *site* www.artmed.com.br), que detalha o F de Welch. Tenham em mente que essas duas técnicas fazem ajustes no F e nos graus de liberdade dos resíduos com o propósito de combater os problemas originados da violação da hipótese da homogeneidade das variâncias.

A pergunta óbvia é qual dos dois procedimentos é melhor. Tomarken e Serlin (1986) revisaram essas e outras técnicas e concluíram que ambas controlam bem a taxa de erro do Tipo I (isto é, quando não existem efeitos na população você, de fato, obtém um F não-significativo). Contudo, em termos de poder (isto é, qual teste é melhor para detectar um efeito quando ele existe), o teste de Welch parece se comportar melhor exceto quando existe uma média extrema com uma grande variância.

Material com direitos autorais

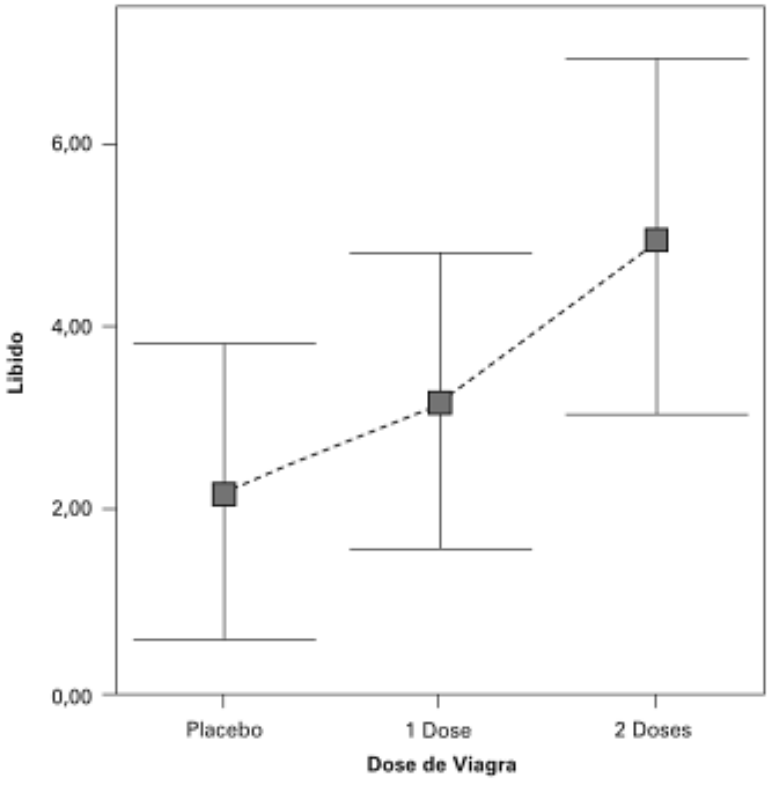


Figura 8.14 Diagrama de barras de erros para os dados do Viagra.

fico será parecido com o da Figura 8.14, mas sem as barras de erros). Escolhi apresentar um diagrama de barras de erro porque ele é mais interessante que apenas o diagrama de linha. Fica claro a partir desse gráfico que todas as barras de erro se sobrepõem, indicando que não existem diferenças entre os grupos (embora essa medida seja aproximada). A linha que une as médias indica uma tendência linear no sentido de que à medida que a dose de Viagra aumenta, cresce também proporcionalmente o nível médio da libido.

A Saída do SPSS 8.3 mostra a tabela das estatísticas descritivas para o procedimento de um fator dos dados do Viagra. Observe que as médias e os desvios padrão correspondem àqueles mostrados na Tabela 8.1. Além disso, o erro padrão é apresentado. Você deve lembrar que o erro padrão é o desvio padrão da distribuição amostral desses dados (assim, para o grupo-placebo, se você retirar muitas amostras da população da qual esses dados vieram, as médias dessas amostras terão um desvio padrão de 0,5831). Intervalos de confiança para as médias também são apresentados. Você já deve estar familiarizado com o que um intervalo de confiança nos informa: se tomarmos 100 amostras da população da qual o grupo-placebo foi retirado e construirmos intervalos de confiança para as médias dessas amostras, 95 desses intervalos irão conter a média da população. Em outras palavras, é provável que a média da população esteja entre 0,5811 e 3,8189. Embora esses diagnósticos não sejam importantes agora, iremos nos referir a eles no decorrer da análise.

A próxima parte da saída é uma tabela que resume o teste de Levene (Saída do SPSS 8.4). Esse teste é projetado para verificar a hipótese nula de que as variâncias dos grupos são iguais. Ele é uma ANOVA conduzida nas diferenças

Material com direitos autorais

de tendência. Essa análise divide o efeito experimental para ver se ele pode ser explicado por um relacionamento linear ou quadrático. Primeiro, vamos tentar uma componente linear. Essa comparação testa se as médias crescem com os grupos de uma forma linear. Novamente a soma dos quadrados e a média dos quadrados são fornecidas, mas o mais importante a ser notado é o valor da razão F e o valor correspondente da significância. Para a tendência linear a razão F é 9,97 e esse valor é significativo a um nível de 0,008. Portanto, podemos dizer que à medida que a dose de Viagra aumenta de nenhum Viagra para uma alta dose do medicamento, a libido cresce proporcionalmente. Na tendência quadrática, essa comparação é testar se o padrão das médias é curvilíneo (isto é, representado por uma curva que tem uma curvatura). O gráfico de barras de erros dos dados recomenda que as médias não podem ser representadas por uma curva e os resultados para a tendência quadrática confirmam isso. A razão F para a tendência quadrática não é significativa (de fato, o valor de F é menor do que 1, o que imediatamente indica que esse contraste não será significativo).

Finalmente, a Saída do SPSS 8.6 mostra as razões F de Welch e de Brow-Forsythe. Não precisamos delas, pois o teste de Levene não foi significativo, indicando que as variâncias podem ser consideradas iguais. Contudo, quando a homogeneidade da variância for violada você deve levar em conta esses valores *em vez* daqueles na tabela principal. Se você estiver interessado em como esses valores são calculados, verifique o Quadro 8.3 e o arquivo **WelchF.pdf** no *site* www.artmed.com.br. No entanto, sendo honesto, isso é bastante confuso e eu recomendo que você fique com os valores calculados na Saída 8.6 do SPSS e acredite que eles fazem o que devem fazer (você também deve notar que os graus de liberdade do erro foram ajustados e precisa lembrar-se disso quando for relatar esses valores!).

Saída do SPSS 8.6

***Robust Tests of Equality of Means* (Testes robustos de igualdade de médias)**

Libido	*Statistic*[a] (Estatística)	*df1* (gl1)	*df2* (gl2)	*Sig.* (Sig.)
Welch	4.320	2	7.943	0.054
Brown-Forsythe	5.119	2	11.574	0.026

a. Assymptotically F distributed (F assintoticamente distribuído)

8.4.2 Saídas para as comparações planejadas ②

Na Seção 8.31, solicitamos que o SPSS realizasse duas comparações planejadas: uma para testar se o grupo-controle era diferente dos dois grupos que receberam o Viagra e outro para ver se as duas doses de Viagra faziam diferença na libido. A Saída do SPSS 8.7 mostra os resultados das comparações que foram solicitadas para os dados do Viagra. A primeira tabela apresenta os contrastes dos coeficientes; esses valores são os que foram digitados na Seção 8.3.1 e vale a pena olhar essa tabela para verificar novamente que os contrastes estão comparando aquilo que eles deveriam comparar! Como uma regra prática rápida, lembre que quando realizamos comparações planejadas arranjamos os pesos de tal forma que comparamos qualquer grupo com um peso positivo contra qualquer grupo com pesos negativos. Portanto, a tabela dos pesos mostra que o contraste 1 compara o grupo-placebo contra os dois grupos-experimentais e o contraste 2 compara o grupo de baixa dosagem contra o de alta dosagem. É útil dar uma olhada nessa tabela para se certificar de que os pesos digitados no SPSS correspondem aos pesos que nós, de fato, pretendíamos informar ao SPSS!

A segunda tabela fornece a estatística para cada contraste. Note que as estatísticas são produzidas para situações nas quais as variâncias dos grupos são iguais e para quando elas são diferentes. Se o teste de Levene foi significativo, você deve ler a parte da tabela rotulada como: **Does not assume equal variances** (Não assume a igualdade de variâncias). Contudo, para esses dados o teste de Levene não foi significativo, portanto, podemos utilizar a parte da tabela denominada **Assume equal variances** (Assume a igualdade de variâncias). A tabela nos informa o valor do contraste, que é a soma ponderada da média dos grupos. Esse valor é obtido toman-

Material com direitos autorais

Saída do SPSS 8.7

Contrast Coefficients **(Coeficientes dos contrastes)**

Contrast (Contraste)o	*Placebo* (Placebo)	*Low Dose* (Baixa Dosgaem)	*High Dose* (Alta Dosagem)
1	−2	1	1
2	0	−1	1

Contrast Tests **(Testes dos contrastes)**

		Contrast (Contraste)	*Value of Contrast* (Valor do contraste)	*Std Error* (Erro padrão)	*t*	*df* (gl)	*Sig. (2-tailed)* (Sig. bilateral)
Libido	*Assume equal variances* (Assume a igualdade de variâncias)	1	3.8000	1.5362	2.474	12	0.029
		2	1.8000	0.8869	2.029	12	0.065
	Does not assume equal variances (Não assume a igualdade de variâncias)	1	3.8000	1.4832	2.562	8.740	0.031
		2	1.800	0.9165	1.964	7.720	0.086

do a média de cada grupo, multiplicando-a pelo peso do contraste de interesse e então somando todos os valores.[6] A tabela também fornece o erro padrão de cada contraste e a estatística t. A estatística *t* é obtida dividindo o valor do contraste pelo erro padrão ($t = 3{,}8/1{,}5362 = 2{,}47$) e isso é comparado com o valor crítico da distribuição *t*. A significância do contraste é dada na coluna final e esse valor é bilateral. Utilizando o primeiro contraste como exemplo, se o tivéssemos utilizado para testar a hipótese geral de que os grupos-experimentais iriam diferir do placebo, deveríamos utilizar esse valor bilateral. Contudo, na realidade testamos a hipótese de que os grupos-experimentais irão aumentar a libido acima dos níveis observados no grupo-placebo: essa hipótese é unilateral. Obtidas as médias para esses grupos e considerando as hipóteses, podemos dividir os valores das significâncias por 2 a fim de obter as probabilidades unilaterais. Assim, para o contraste 1, podemos dizer que tomar Viagra aumenta significativamente a libido comparado ao grupo-controle ($p = 0{,}0145$). Para o contraste 2 temos novamente uma hipótese unilateral (de que uma alta dosagem de Viagra irá aumentar a libido mais do que uma baixa dosagem) e as médias suportam essa hipótese. A significância para o contraste 2 nos informa que uma alta dose de Viagra aumenta significativamente a libido mais do que uma baixa dosagem ($p_{\text{unilateral}} = 0{,}065/2 = 0{,}0325$). Note que se não tivéssemos uma hipótese específica sobre qual grupo teria a média mais alta, deveríamos concluir que a dose de Viagra não teria um efeito significativo na libido. Por isso, é importante que geremos as hipóteses antes de coletar qualquer dado porque esse método de descoberta científica é mais poderoso.

Em resumo, existe um efeito global do Viagra na libido. As comparações planejadas revelaram que o Viagra aumentou significativamente a libido quando comparado ao grupo-controle, $t(12) = 2{,}47$, $p < 0{,}05$, e que uma alta dosagem aumentou significativamente a libido quando comparada com uma baixa dosagem, $t(12) = 2{,}03$, $p < 0{,}05$.

8.4.3 Saídas para os testes *post hoc* ②

Se não tivéssemos hipóteses específicas sobre o efeito que o Viagra exerce na libido, poderíamos executar testes *post hoc* para comparar todos os grupos de participantes entre eles. De fato, solicitamos ao SPSS que fizesse

[6] Para o primeiro contraste, esse valor é:

$$\sum(\bar{X}W) = [(2{,}2 \times -2) + (3{,}2 \times 1) + (5{,}0 \times 1)] = 3{,}8$$

Material com direitos autorais

isso (veja a Seção 8.3.2), e os resultados dessa análise são mostrados na Saída 8.8 do SPSS. Essa tabela mostra os resultados do teste de Tukey (conhecido como Tukey HSD)[7], do procedimento de Games-Howell e o teste de Dunnet, que foram especificados anteriormente. Se observarmos primeiro o teste de Tukey (porque não temos razão para duvidar que as variâncias populacionais sejam diferentes), fica claro a partir da tabela que cada grupo de participantes é comparado aos outros grupos restantes. Para cada par de grupos, a diferença entre as médias dos grupos é apresentada bem como o erro padrão, o nível de significância e um intervalo de 95% de confiança para essa diferença. Primeiro, o grupo-placebo é comparado ao grupo de baixa dosagem e revela uma diferença não-significativa (*Sig.* é maior do que 5%), mas quando comparado ao grupo de alta dosagem existe uma diferença significativa (*Sig.* é menor do que 5%). Essa descoberta é interessante porque nossa comparação planejada mostrou que qualquer dose de Viagra produz um aumento significativo na libido, embora essas comparações indiquem que uma dosagem baixa não faça

[7] A sigla HSD signfica *Honestily Significant Difference* (Diferença Honestamente Significativa).

Saída do SPSS 8.8

***Multiple Comparisons* (Comparações múltiplas)**

Dependent Variable: Libido (Variável dependente: libido)

	(I) Dose of Viagra (I Dose de Viagra)	*(J) Dose of Viagra* (J Dose de Viagra)	*Mean Difference (I-J)* (Diferença Média I-J)	*Std. Error* (Erro Padrão)	*Sig.* (Sig.)	*95% Confidence Interval* (Intervalo de 95% Confiança)	
						Lower Bound (Limite Inferior)	*Upper Bound* (Limite Superior)
HSD de Tukey	Placebo	*Low Dose* (Baixa Dosagem)	–1.0000	0.88694	0.516	–3.3662	1.3662
		High Dose (Alta Dosagem)	–2.8000*	0.88694	0.021	–5.1662	–0.4338
	Low Dose (Baixa Dosagem)	Placebo	1.0000	0.88694	0.516	–1.3662	3.3662
		High Dose (Alta Dosagem)	–1.8000	0.88694	0.147	–4.1662	0.5662
	High Dose (Alta Dosagem)	Placebo	2.8000*	0.88694	0.021	0.4338	5.1662
		Low Dose (Baixa Dosagem)	1.8000	0.88694	0.147	–0.5662	4.1662
Games-Howell	Placebo	*Low Dose* (Baixa Dosagem)	–1.0000	0.82462	0.479	–3.3563	1.3563
		High Dose (Alta Dosagem)	–2.8000*	0.91652	0.039	–5.4389	–0.1611
	Low Dose (Baixa Dosagem)	Placebo	1.0000	0.82462	0.0479	–1.3563	3.3563
		High Dose (Alta Dosagem)	–1.8000	0.91652	0.185	–4.4389	0.8389
	High Dose (Alta Dosagem)	Placebo	2.8000*	0.91652	0.039	0.1611	5.4389
		Low Dose (Baixa Dosagem)	1.8000	0.91652	0.185	–0.8389	4.4389
Dunnett *t* (> control)[a] (> controle)	*Low Dose* (Baixa Dosagem)	Placebo	1.0000	0.88694	0.227	0.8697	
	High Dose (Alta Dosagem)	Placebo	2.8000*	0.88694	0.008	0.9303	

* A diferença média é significativa ao nível de 5%.

a O test *t* de Dunnett trata um grupo como controle e compara todos os outros contra esse grupo-controle.

Material com direitos autorais

efeito. Por que existe essa contradição? (Pense um pouco nisso antes de prosseguir.)

Na Seção 8.2.10.2, expliquei que a primeira comparação planejada compara os grupos-experimentais com o grupo-placebo. Especificamente, ela compara a média dos dois grupos-experimentais ((3,2 + 5,0)/2 = 4,1) com a média do grupo-placebo (2,2). Assim, a diferença entre esses dois valores (4,1 – 2,2 = 1,9) é que foi considerada significativa. Nos testes *post hoc*, quando o grupo de baixa dosagem é comparado com o grupo-placebo, o contraste está testando se a diferença entre as médias desses dois grupos é significativa. A diferença, nesse caso, é de apenas 1, comparada com uma diferença de 1,9 para a comparação planejada. Essa explicação ilustra como é possível ter resultados aparentemente contraditórios entre as comparações planejadas e as comparações *post hoc*. Mais importante, ela ilustra o cuidado que devemos ter na interpretação dos contrastes planejados.

O grupo de baixa dosagem é então comparado com o grupo-placebo e com o grupo de alta dosagem. Note primeiro que o contraste envolvendo os grupos-placebo e de baixa dosagem é idêntico àquele recém-descrito. A única informação nova é a comparação entre as duas condições experimentais. As médias dos grupos diferem por 1,8, o que não é significativo. Esse resultado também contradiz as comparações planejadas (lembre que o contraste 2 comparou esses grupos e encontrou uma diferença significativa). Pense por que essa contradição está ocorrendo. Por um lado, testes *post hoc* por sua natureza são bilaterais (você os utiliza quando não tiver uma hipótese e você não prevê a direção da hipótese porque ela não existe!) e o contraste 2 foi significativo somente quando considerado uma hipótese unilateral. Contudo, mesmo com um nível bilateral a comparação planejada foi mais próxima na significância que o teste *post hoc* e esse fato ilustra que os procedimentos *post hoc* são mais conservadores (isto é, apresentam menos poder de detectar verdadeiras diferenças) que as comparações planejadas.

O restante da tabela descreve o teste de Games-Howell e uma rápida inspeção revela o mesmo padrão de resultados: os únicos grupos que diferem significativamente foram o de alta dosagem e o placebo. Esses resultados confirmam nossas conclusões a partir do teste de Tukey porque mesmo se as variâncias populacionais não forem iguais (o que parece improvável uma vez que as variâncias amostrais são bastante semelhantes), o perfil dos resultados ainda será verdadeiro. Finalmente, o teste de Dunnet é descrito e você, espero, irá lembrar que solicitamos que o computador comparasse os dois grupos-experimentais contra o grupo-controle utilizando uma hipótese unilateral de que a média do grupo-controle seria menor do que as médias dos grupos-experimentais. Mesmo com uma hipótese unilateral, os níveis de libido com baixa dosagem não são diferentes do grupo-placebo. Contudo, no grupo de alta dosagem a libido é significativamente maior do que no grupo-placebo.

A tabela na Saída 8.9 do SPSS mostra os resultados do teste de Tukey e do REGWQ. Esses testes mostram subconjuntos de grupos que têm a mesma média. Portanto, o teste de Tukey cria dois subconjuntos de grupos com médias estatisticamente semelhantes. O primeiro desses subconjuntos contém o grupo-placebo e o de baixa dosagem (indicando que esses dois grupos apresentam médias semelhantes) e o segundo desses subconjuntos contém os grupos de alta e baixa dosagem. Esses resultados demonstram que o grupo-placebo tem uma média semelhante ao grupo de baixa dosagem, mas não ao de alta dosagem, e que o grupo de baixa dosagem apresenta média semelhante tanto ao de alta dosagem quanto ao grupo-placebo. Em outras palavras, os únicos grupos que apresentam médias significantemente diferentes são os grupos-placebo e de alta dosagem. O teste fornece um valor da significância para cada subconjunto e fica claro a partir desses valores que os grupos nos subconjuntos não têm médias significativas (como o indicado pelos valores *Sig.* que são maiores do que 0,05).

Esses cálculos utilizam a média harmônica dos tamanhos amostrais. A média harmônica é uma versão ponderada da média

Material com direitos autorais

Saída do SPSS 8.9

Libido

Dose of Viagra (Dose de Viagra)		*N*	*Subset for alpha* = .05 (Subconjunto para alfa = 0,05)	
			1	*2*
Tukey HSD[a] (HSD de Tukey)	Placebo	5	2.2000	
	Low Dose (Baixa Dosagem)	5	3.2000	3.2000
	High Dose (Alta Dosagem)	5		5.0000
	Sig.		0.516	0.147
Ryan-Einot-Gabriel-Welsch Range	Placebo	5	2.2000	
	Low Dose (Baixa Dosagem)	5	3.2000	3.2000
	High Dose (Alta Dosagem)	5		5.0000
	Sig.		0.282	0.065

Médias para subconjuntos homogêneos são apresentadas.
a Utiliza a média harmônica dos tamanhos amostrais = 5.

Dica da Samanta Ferrinho

- A ANOVA independente de um fator compara várias médias, quando essas médias vêm de grupos diferentes de pessoas; por exemplo, se você tem várias condições experimentais e utiliza diferentes participantes em cada condição.
- Quando você tiver gerado hipóteses específicas antes do experimento ser realizado utilize as *comparações planejadas*, mas se você não tiver hipóteses específicas, utilize os testes *post hoc*.
- Existem vários testes *post hoc* diferentes: quando você tiver amostras do mesmo tamanho e variâncias homogêneas utilize o REGWQ ou o HSD de Tukey. Se os tamanhos das amostras não diferem muito, utilize o procedimento de Gabriel, mas se os tamanhos amostrais são bem diferentes, use o GT2 de Hochberg. Se existir qualquer dúvida sobre a homogeneidade das variâncias, utilize o procedimento de Games-Howell.
- Teste a homogeneidade das variâncias utilizando o teste de Levene. Encontre a tabela com esse rótulo: se o valor na coluna denominada *Sig.* é menor do que 0,05, a hipótese foi violada. Nesse caso, vá para a tabela denominada *Robust Tests of Equality of Means* (Testes robustos de igualdade de médias). Se a homogeneidade das variâncias não foi violada (isto é, a significância do teste de Levene é maior do que 5%), vá para a tabela denominada ANOVA.
- Na tabela denominada ANOVA (ou Testes Robustos para a Igualdade de Médias – veja acima), olhe para a coluna *Sig.:* se o valor for menor do que 0,05, as médias dos grupos são significativamente diferentes.
- Para contrastes e testes *post hoc*, olhe novamente a coluna *Sig.* para descobrir se suas comparações são significativas (elas serão se o valor da significância for menor do que 0,05).

que leva em conta o relacionamento entre variâncias e tamanhos amostrais (veja Howell, 2002, p. 234). Embora você não precise saber os caminhos intrincados da média harmônica, é útil que o tamanho amostral harmônico seja utilizado, porque elimina o viés que poderia ser introduzido pelos tamanhos amostrais diferenciados. Assim, a ANOVA de um fator fornece resultados confiáveis mesmo com delineamentos desbalanceados (embora os delineamentos balanceados também apresentem vantagens!).

Material com direitos autorais

8.8 O QUE DESCOBRIMOS SOBRE ESTATÍSTICA? ①

Neste capítulo abordamos a análise de variância (ANOVA), que é o tópico dos próximos capítulos também. A ANOVA independente de um fator é utilizada em situações onde queremos comparar várias médias e os dados foram coletados utilizando diferentes participantes em cada condição. Comecei explicando que se fizermos muitos testes *t* nos mesmos dados, a nossa taxa de erro do Tipo I irá inflacionar. Portanto, utilizamos a ANOVA em vez dos testes t. Relatei como a ANOVA pode ser conceitualizada como um MLG (Modelo Linear Geral), assim ela é, de fato, o mesmo que uma regressão múltipla. Como em uma regressão múltipla, existem três médias importantes que utilizamos na ANOVA: a soma total dos quadrados, SS_T (uma medida da variabilidade nos dados), a soma dos quadrados do modelo, SS_M (uma medida de quanta variabilidade pode ser explicada pela nossa manipulação experimental) e SS_R (uma medida de quanta variabilidade não pode ser explicada pela manipulação experimental). Descobrimos que a razão F é simplesmente a razão entre a variância que podemos explicar pela variância que não podemos. Também descobrimos que uma razão F significativa nos informa apenas que nossos grupos diferem, mas não como eles diferem. Para descobrir onde estão as diferenças temos duas opções: determinar contrastes específicos para testar hipóteses (contrastes planejados) ou testar cada grupo contra cada um dos demais grupos (*testes post hoc*). A primeira é utilizada quando temos hipóteses formuladas antes da condição do experimento e a segunda é para explorar os dados quando nenhuma hipótese foi feita. Finalmente, descobrimos como programar esses procedimentos no SPSS.

8.9 TERMOS-CHAVE QUE DESCOBRIMOS

- Análise de Variância (ANOVA)
- Correção de Bonferroni
- F de Brown-Forsythe
- Tendência cúbica
- Graus de liberdade
- Desvios dos contrastes
- Contraste de Helmert
- ANOVA independente
- Ômega ao quadrado
- Ortogonal
- Comparações pareadas
- Contrastes planejados
- Contrastes polinomiais
- Contraste diferença (Contraste reverso de Helmert)
- Taxa de erro experimental
- Taxa de erro de conjunto
- Variância geral
- Média harmônica
- Tendência Quadrática
- Tendência Quártica
- Contraste repetido
- Contraste simples
- Ponderações (pesos)
- F de Welch
- Testes *post hoc*

8.10 TAREFAS DO ALEX ESPERTO

- **Tarefa 1:** Imagine que eu esteja interessado em como diferentes métodos de ensino afetam o conhecimento dos estudantes. Notei que alguns professores são arrogantes e distantes no seu estilo de ensinar e humilham qualquer um que se atrever a formular uma questão, enquanto outros estimulam e apoiam perguntas e comentários. Lecionei o mesmo conteúdo em três cursos de estatística. Para um grupo eu andava pela sala com uma grande vara e batia em qualquer um que fizesse uma pergunta idiota ou respondesse errado (*punitivo*). No segundo grupo utilizei meu método normal de ensino que é incentivar o aluno a discutir assuntos que eles achassem difíceis e fornecendo a qualquer um que se esforçasse uma recompensa (*incentivador*). No último grupo permaneci indiferente e não recompensei nem tampouco puni o esforço

Material com direitos autorais

(*indiferente*). Usei a nota nos alunos no exame (em percentual) como uma medida dependente. Com base nas teorias do condicionamento operante, espera-se que o método punitivo seja a forma malsucedida de reforço de aprendizagem, mas espera-se que o método incentivador seja bem-sucedido. Dessa forma, uma previsão é que o método incentivador irá produzir a melhor aprendizagem. Uma segunda hipótese é que o método punitivo irá retardar a aprendizagem de modo que ele é pior do que a abordagem indiferente. Os dados estão no arquivo **Teach.sav**. Execute uma ANOVA de um fator e utilize as comparações planejadas para testar as hipóteses que (1) o método incentivador resulta em melhores notas do que o método punitivo e o indiferente; e (2) o método indiferente irá produzir resultados significativamente melhores do que o punitivo. ②

- **Tarefa 2**. No Capítulo 13 (Seção 13.4) existem alguns dados sobre se comer refeições com soja reduz a contagem de espermatozoides. Dê uma olhada nessa seção, acesse os dados para esse exemplo, mas utilize a ANOVA para analisá-los. Qual é a diferença sobre o que você encontrou e o que está na Seção 13.4? Por que você acha que essa diferença surgiu? ②
- **Tarefa 3**. Estudantes (e professores) adoram os seus celulares, o que é bastante preocupante dado as recentes controvérsias sobre a relação entre o uso de celulares e tumores cerebrais. A ideia básica é que telefones celulares emitem micro-ondas e, assim, manter um próximo à cabeça por um número elevado de horas é mais ou menos como colocar o cérebro num forno de micro-ondas e apertar o botão "cozinhar até que fique bem passado". Se quiséssemos testar isso de forma experimental, poderíamos arranjar seis grupos de pessoas e fixar um telefone móvel na suas cabeças (de forma que eles não possam remover). Então, por controle remoto, ligaríamos os telefones por certo tempo cada dia. Após seis meses, mediríamos o tamanho de qualquer tumor (em mm^3) próximo ao local da antena (logo atrás da orelha). Os seis grupos apresentariam 0, 1, 2, 3, 4 ou 5 horas diárias de micro-ondas do telefone por seis meses. Os dados desse exemplo estão no arquivo **Tumour.sav** (De Fiel e Hole, 2003, existe uma solução bastante detalhada no arquivo). ②

As respostas estão no arquivo **Answers (Chapter 8).pdf** no *site* www.artmed.com.br.

8.11 LEITURAS COMPLEMENTARES

HOVELL, D. C. *Statistical methods for psychology.* Belmont (CA): Duxbury, 2002. 5nd ed. Os Capítulos 11 e 12 fornecem uma abordagem detalhada da ANOVA e o Capítulo 16 cobre os MLG (Modelos Lineares Generalizados).

IVERSEN, G. R., NORPOTH, H. *ANOVA.* Sage university paper series on quantitative applications in the social sciences, 07-001. Newbury Park (CA): Sage. 1987. 2nd ed. Texto de alto nível, mas uma boa leitura para aqueles que gostam de matemática.

KLOCKARS, A. J., SAX, G. *Multiple comparisons.* Sage university paper series on quantitative applications in the social sciences, 07-061. Newbury Park (CA): Sage, 1986. De alto nível, mas com uma cobertura completa sobre comparações múltiplas – para mim esse livro é melhor do que o de Toothaker para comparações planejadas.

ROSENTHAL, R. ROSNOW, R. L, RUBIN, D. B. *Contrasts and effect sizes in behavioral research: a correlational approach.* Cambridge: Cambridge University Press, 2000. Livro fantástico sobre comparações planejadas por três dos melhores escritores sobre estatística.

ROSNOW, R. L, ROSENTHAL, R. *Beginning behavioral research: a conceptual primer.* Englewood Cliffs (NJ): Hall, 2005. 5ª ed. E eles escreveram outro ótimo livro!

TOOTHAKER, L. E. *Multiple comparison procedures.* Sage university paper series on quantitative applications in the social sciences, 07-089. Newbury Park (CA): Sage, 1993.

Material com direitos autorais

Também de alto nível, mas fornece uma excelente abordagem sobre os procedimentos *post hoc*.

WRIGHT, D. B. *First steps in statistics*. London: Sage, 2002. Se esse capítulo for muito complicado, o Capítulo 7 do Wright fornece uma introdução básica à ANOVA de fácil leitura.

Material com direitos autorais

9

ANÁLISE DE COVARIÂNCIA, ANCOVA (MLG 2)

9.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ②

No capítulo anterior, vimos como a ANOVA de um fator pode ser caracterizada em termos de uma equação de regressão múltipla que utiliza variáveis *dummy* (auxiliares) para codificar a pertinência aos grupos. Além disso, no Capítulo 5 vimos como a regressão múltipla pode incorporar diversas variáveis previsoras contínuas. Dessa forma, não deve ser surpresa verificar que a equação de regressão para ANOVA pode ser estendida para incluir uma ou mais variáveis contínuas que podem prever a saída (ou variável dependente). Variáveis contínuas como essas – que não são parte da manipulação experimental principal, mas que têm influência na variável dependente – são conhecidas como *covariáveis* e podem ser incluídas em uma ANOVA. Quando mensuramos as covariáveis e as incluímos na ANOVA, denominamos o procedimento de **análise de covariância** (ou ANCOVA). Este capítulo está centrado nessa técnica. Começaremos apresentando a teoria da ANCOVA de uma forma breve (não se preocupe, não entrarei em muitos detalhes) e uma hipótese adicional, a homogeneidade dos coeficientes de regressão. Também iremos mostrar um exemplo com a ajuda do SPSS e ver como interpretar e relatar os resultados dessa técnica.

9.2 O QUE É ANCOVA? ②

No último capítulo, usamos um exemplo que verificava os efeitos do Viagra na libido. Vamos pensar em outras coisas além do Viagra que podem influenciar a libido: bem, a mais óbvia é a libido do parceiro sexual do participante (são necessários dois para dançar um tango!), mas existem outras coisas, como medicamentos que podem inibir a libido (antidepressivos ou a pílula anticoncepcional) e a fadiga. Se essas variáveis (chamadas de **covariáveis**) são medidas, é possível controlar a influência que elas exercem na variável dependente pela sua inclusão no modelo de regressão. Do que aprendemos sobre regressão hierárquica (veja o Capítulo 5), deve ter ficado claro que se entrarmos com a covariável no modelo de regressão primeiro e depois entrarmos com as variáveis *dummies* (auxiliares) representando a manipulação experimental, será possível ver qual o efeito que uma

Material com direitos autorais

variável independente tem *após* o efeito da covariável. Assim, nós controlamos o efeito da covariável. Existem dois motivos para incluir covariáveis na ANOVA.

- **Para reduzir a variância do erro dentro do grupo:** Na discussão sobre a ANOVA e o teste *t* nos acostumamos com a ideia de que avaliamos o efeito de um experimento comparando o montante de variabilidade nos dados que o experimento pode explicar contra a variabilidade que ele não consegue explicar. Se conseguirmos explicar parte dessa variância "inexplicada" (SS_R) em termos de outras variáveis (covariáveis), reduziremos a variância do erro e determinaremos com maior precisão o efeito da variável independente (SS_M).
- **Eliminação dos confundidores:** Em qualquer experimento pode haver variáveis não medidas que confundem os resultados (isto é, variáveis que podem variar sistematicamente com a manipulação experimental). Se for sabido que uma variável pode influir na variável dependente sendo medida, a ANCOVA é ideal para remover a parcialidade dessa influência. Uma vez que uma possível variável de confusão foi identificada, ela pode ser medida e entrar na análise como uma covariável.

Existem outras razões para incluir covariáveis na ANOVA, mas como não pretendo descrever os cálculos da ANCOVA, recomendo que o leitor interessado consulte Wildt e Ahtola (1978) ou Stevens (1992, Capítulo 9).

Imagine que o pesquisador que conduziu o estudo sobre o Viagra no capítulo anterior subitamente se deu conta de que a libido do parceiro do participante poderia afetar a libido do próprio participante (especialmente porque a medida da libido foi comportamental). Ele repetiu o estudo num conjunto diferente de participantes, mas dessa vez medindo a libido do parceiro. A libido do parceiro foi medida levando-se em conta a frequência com que eles tentaram iniciar um contato sexual. No capítulo anterior, vimos que esse cenário experimental pode ser caracterizado em termos de equação (8.2). Pense no que aprendemos sobre a regressão múltipla (Capítulo 5) e você verá que essa equação pode ser estendida para incluir essa covariável (veja a equação 9.1):

$$\begin{aligned}\text{Libido}_i &= b_0 + b_3\text{Covariável}_i + b_2\text{Alto}_i \\ &+ b_1\text{Baixo}_i + \varepsilon_i \\ \text{Libido}_i &= b_0 + b_3\text{Libido do Parceiro}_i \\ &+ b_2\text{Alto}_i + b_1\text{Baixo}_i + \varepsilon_i \end{aligned} \quad (9.1)$$

9.3 EXECUTANDO A ANCOVA NO SPSS ②

Os dados para esse exemplo estão na Tabela 9.1 e podem ser encontrados no arquivo **ViagraCovariate.sav**. A Tabela 9.1 mostra a libido dos participantes, dos parceiros e as médias (os desvios padrão estão entre parênteses) dos vários escores. Recomendo incluir esses dados no editor manualmente. Faça isso da mesma forma que foi feito para os dados do Viagra do capítulo anterior, exceto que uma variável a mais deve ser criada para incluir os valores da covariável.

9.3.1 Atribuindo dados ①

Os dados devem ser exibidos no editor de dados assim como eles estão na Tabela 9.1 (excluindo as linhas das médias e dos desvios padrão). Para tanto, crie uma variável código chamada de **dose** e use a opção **Labels** (Rótulos) para definir o valor dos rótulos (como no Capítulo 8, recomendo 1 = placebo, 2 = dose baixa, 3 = dose alta). Existem diferentes números de participantes em cada condição, assim, você precisa incluir nove valores de 1 nessa coluna (de modo que as primeiras nove linhas contenham o valor 1), seguido por oito linhas contendo o valor 2 e mais 14 linhas contendo o valor 3. Nesse ponto, você deve ter uma coluna com 30 linhas de dados. A seguir, crie uma segunda variável chamada **libido** e entre com os 30 escores que correspondem à libido de cada pessoa. Finalmente, crie uma terceira variável chamada de **partner (parceiro)** e use a opção **Labels** (Rótulos) para dar a essa variável o título de "*Partners libido*"

Material com direitos autorais

Figura 9.3 Caixa de diálogo **Options** (Opções) para o MLG univariado.

terações de fator) e transfira-a para o quadro denominado **Display Means** (Mostrar as Médias) clicando em ▶. Uma vez que a variável foi transferida, a caixa rotulada de **Compare main effects** (Comparar efeitos principais) torna-se ativa e você deve selecionar essa opção (Compare main effects). Se essa opção é selecionada, o quadro denominado **Confidence interval adjustment** (Ajuste do intervalo de confiança) torna-se ativo e você pode clicar em ▼ para escolher uma opção entre três possíveis níveis de ajuste. O padrão é não ter ajustamentos. A primeira opção é realizar o procedimento *post hoc* Tukey LSD (que não recomendo); a segunda é solicitar uma correção de Bonferroni (essa eu recomendo) e a opção final é uma correção de Sidak. A correção de Sidak é parecida com a correção de Bonferroni, mas menos conservadora; desse modo, deve ser selecionada se você está preocupado com a perda de poder associada aos valores corrigidos de Bonferroni. Para esse exemplo utilize a correção de Sidak (iremos usar a de Bonferroni mais adiante neste capítulo). Assim como a realização dos testes *post hoc* para a variável **dose** (dose), colocar a variável **dose** (dose) no quadro **Display Means for** (Mostrar as Médias para) irá resultar numa tabela de médias marginais estimadas para essa variável. Essas médias fornecem uma estimativa das médias do grupo ajustadas (isto é, as médias após a contribuição da covariável). Quando você tiver selecionado as opções solicitadas (veja o Quadro 9.1), clique em Continue (continuar) para retornar à caixa de diálogo principal. Existem outras opções disponíveis na caixa de diálogo principal. Por exemplo, se você tem diversas variáveis independentes, pode representá-las graficamente umas contra as outras (o que é útil para interpretar efeitos de interação – veja a Seção 10.3.3). Para essa análise existe somente uma variável independente, assim, podemos clicar em OK a fim de executar a análise.

9.4 INTERPRETANDO AS SAÍDAS DA ANCOVA ②

9.4.1 Análise principal ②

A Saída 9.1 do SPSS mostra (como ilustração) a tabela da ANOVA para esses dados quando a covariável não está incluída. Fica claro a partir do valor da significância, maior do que 0,05, que não há diferenças na libido

Material com direitos autorais

Saída 9.1 do SPSS

***Tests of Between-Subjects Effects* (Testes de efeitos entre participantes)**

Dependent Variable: Libido (Variável Dependente: Libido)

Source (Fonte)	*Type III Sum of Squares* (Soma dos quadrados do tipo III)	*df* (gl)	*Mean Square* (Média ao quadrado)	*F*	*Sig.*
Corrected Model (Modelo Corrigido)	16.844[a]	2	8.422	2.416	0.108
Intercept (Intercepto)	535.184	1	535.184	153.533	0.000
DOSE (Dose)	16.844	2	8.422	2.416	0.108
Error (Erro)	94.123	27	3.486		
Total (Total)	683.000	30			
Corrected Total (Total corrigido)	110.967	29			

a. R Squared = .152 (Adjusted R Squared = .089 (R ao quadrado = 0,152 (R ao quadrado ajustado = 0,089))

foi violada). Entretanto, como foi mencionado na Seção 3.6, o teste de Levene não é necessariamente a melhor maneira de julgar se as variâncias são desiguais o suficiente para causar problemas. Uma boa maneira de se assegurar disso é verificar novamente os valores da maior e da menor variância. Para os nossos três grupos temos os desvios padrão de 1,79 (grupo-placebo), 1,46 (grupo de baixa dosagem) e 2,12 (grupo de alta dosagem) – veja a Tabela 9.1. Se elevarmos ao quadrado esses valores, teremos variâncias de: 3,20 (grupo-placebo), 2,13 (grupo de baixa dosagem) e 4,49 (grupo de alta dosagem). Podemos, então, dividir a maior variância pela menor: nesse caso, 4,49/2,13 = 2,11. Se o valor resultante for menor do que 2, provavelmente não precisaremos nos preocupar muito; se for maior do que 2 (como é o caso aqui), provavelmente teremos problemas! Entretanto, não vamos nos preocupar muito com as diferenças entre as variâncias por enquanto.

O formato dessa tabela da ANOVA é, em grande parte, o mesmo que a da tabela sem a covariável, exceto que agora temos uma linha adicional com informações sobre a covariável (**partner** (parceiro)). Observando os valores das significâncias, fica claro que a covariável prevê significativamente a variável dependente porque esse valor é menor do que 0,05. Portanto, a libido de um participante é influenciada pela libido do parceiro. O mais interessante é que quando o efeito da libido do parceiro é removido, o efeito do Viagra se torna significante (p é 0,016, o que é menor do que 0,05). A quantidade de variação de responsabilidade do modelo (SS_M) aumentou para 34,75 unidades (*corrected model* (modelo corrigido)), da qual o Viagra é responsável por 28,34 unidades. Ainda mais importante, a grande variação na libido que é de responsabilidade da covariável significa que a variância não-explicada (SS_R) foi reduzida para 76,22 unidades. Note que o SS_T não mudou; o que mudou foi como a variação total é explicada.

Esse exemplo ilustra como a ANCOVA pode auxiliar a exercer um controle experimental mais rígido, levando em consideração as variáveis de confusão, para nos dar uma medida "mais pura" do efeito da manipulação experimental. Sem levar em consideração a libido dos parceiros participantes, podemos concluir que o Viagra não teve efeito na libido, mas, na verdade, ele teve. Entretanto, o efeito da libido do parceiro parece mais forte do que o do Viagra. Retomando as médias do grupo da Tabela 9.1 para os dados da libido, fica claro que a significância da ANOVA reflete a diferença entre o grupo-placebo e os dois grupos-experimentais (porque os grupos de doses alta e baixa têm médias semelhantes – 4,88 e 4,85 – enquanto a média do grupo-placebo é bem mais baixa, 3,22). Contudo, será necessário realizar alguns contrastes para verificar isso.

Material com direitos autorais

Esses contrastes são consistentes com o que foi especificado: todos os grupos são comparados com o primeiro grupo. As diferenças dos grupos são apresentadas: o valor da diferença, o erro padrão, o valor da significância e o intervalo de 95% de confiança. Esses resultados mostram que o grupo de baixa dosagem não teve uma libido significativamente diferente do que a do grupo-placebo (contraste 1, $p = 0{,}804$), mas que o grupo da alta dosagem diferiu significativamente do grupo-placebo ($p = 0{,}005$). Esses resultados são consistentes com as estimativas de regressão paramétrica (na verdade, note que o contraste 2 é idêntico aos parâmetros da regressão para DOSE = 1 da seção anterior).

Novamente, tudo isso parece muito estranho: o efeito significativo da libido parece refletir uma diferença entre o grupo-placebo e os dois grupos do Viagra (que apresentam médias com valores semelhantes), mas os contrastes até agora contradizem essas conclusões. A razão para essa inconsistência é que a conclusão inicial foi baseada nas médias dos grupos que não tinham sido ajustadas para o efeito da covariável. Esses valores nada informam sobre as diferenças dos grupos refletidas na significância da ANCOVA. A Saída 9.5 do SPSS fornece os valores ajustados das médias dos grupos e são esses os valores que devem ser usados para a interpretação (essa é a razão principal para selecionar a opção **Display Means for** (Mostre as Médias para)). As médias ajustadas mostram um padrão bem diferente de respostas: parece que a significância da ANCOVA reflete a diferença entre o grupo de alta dosagem e os grupos de baixa dosagem e placebo. Os grupos de baixa dosagem e placebo parecem ter médias ajustadas bem similares, indicando que uma dosagem baixa de Viagra não aumenta a libido acima de níveis normais, sendo necessário para tanto uma alta dosagem! Essas conclusões sustentam o que sabemos dos contrastes e dos parâmetros da regressão, mas podem ser verificadas com os testes *post hoc* especificados no menu **options** (opções).

A Saída 9.6 do SPSS mostra os resultados das comparações *post hoc* corrigidas pelo método de Sidak que foram solicitadas como parte da caixa de diálogo **options** (opções). A diferença significativa entre o grupo de alta dosagem e o placebo permanece, mas a diferença significativa entre os grupos de alta e baixa dosagem pelos parâmetros da regressão (Saída 9.3 do SPSS) desapareceu (p é somente 0,14). Essa contradição pode ser resultado de uma perda de poder nos testes *post hoc* (lembre que as comparações planejadas têm mais poder para detectar efeitos do que procedimentos *post hoc*). Entretanto, poderia haver outras razões para essas comparações serem não-significativas e devemos ter cautela nas nossas interpretações da significância da ANCOVA e das comparações subsequentes.

Saída 9.5 do SPSS

***Estimates* (Estimativas)**

Dependent Variable: Libido (Variável dependente: libido)

Dose of Viagra (Dose de Viagra)	*Mean* (Média)	*Std. Error* (Erro Padrão)	*95% Confidence Interval* (Intervalo de 95% de confiança)	
			Lower Bound (Limite Inferior)	*Upper Bound* (Limite Superior)
Placebo (Placebo)	3.313[a]	0.572	2.138	4.489
Low Dose (Baixa Dosagem)	3.027[a]	0.962	1.049	5.004
High Dose (Alta Dosagem)	5.920[a]	0.644	4.597	7.244

a. As covariáveis que aparecem no modelo são avaliadas nos seguintes valores: Libido do Parceiro = 4,30.

9.4.3 Interpretando as covariáveis ②

Já mencionei que as estimativas dos parâmetros (Saída 9.3 do SPSS) nos mostram como interpretar a covariável. Se o valor de b para a covariável for positivo, isso significa que a covariável e a variável de saída tem um relacionamento positivo (à medida que a covariável aumenta, o valor de saída também aumenta). Se o valor de b for negativo significa o oposto: a covariável e a variável

Material com direitos autorais

Saída 9.6 do SPSS

***Pairwise Comparisons* (Comparações aos pares)**

Dependent Variable: Libido (Variável dependente: libido)

(I) Dose of Viagra (I) Dose de Viagra	*(J) Dose of Viagra* (J) Dose de Viagra	*Mean Difference* (Diferença Média)(I – J)	*Std. Error* (Erro Padrão)	*Sig.*[a]	*95% Confidence Interval for Difference*[a] (Intervalo de 95% de confiança para a diferença)	
					Lower Bound (Limite inferior)	*Upper Bound* (Limite Superior)
Placebo (Placebo)	*Low Dose* (Baixa Dosagem)	0.287	1.144	0.992	–2.631	3.205
	High Dose (Alta Dosagem)	–2.607*	0.842	0.014	–4.756	–0.458
Low Dose (Baixa Dosagem)	*Placebo* (Placebo)	– 0.287	1.144	0.992	–3.205	2.631
	High Dose (Alta Dosagem)	–2.894	1.411	0.144	–6.493	0.706
High Dose (Alta Dosagem)	*Placebo* (Placebo)	2.607*	0.842	0.014	0.458	4.756
	Low Dose (Baixa Dosagem)	2.894	1.411	0.144	–0.706	6.493

Baseado nas médias marginais estimadas.
* A diferença média é significativa ao nível de 0,05.
a. Ajuste para as comparações múltiplas: Sidak.

de saída têm um relacionamento negativo (à medida que a covariável aumenta, a variável de saída diminui). Para esses dados o valor de *b* foi positivo, indicando que à medida que a libido do parceiro aumenta, também aumenta a do participante. Outra maneira de descobrir a mesma coisa é simplesmente construir um diagrama de dispersão da covariável contra a saída. Aprendemos sobre diagramas de dispersão na Seção 4.4, então dê uma olhada lá para descobrir como se faz um. A Figura 9.4 mostra o diagrama de dispersão resultante para esses dados e confirma o que já sabemos: o efeito da covariável é que à medida que a libido do parceiro aumenta, o mesmo acontece com a libido do participante (como mostra a inclinação da linha da regressão).

9.5 ANCOVA EXECUTADA COMO UMA REGRESSÃO MÚLTIPLA ②

Embora a ANCOVA esteja basicamente concluída, talvez seja interessante refazer a análise como uma regressão múltipla hierárquica. Como um exercício, entre esses dados e conduza uma análise você mesmo acrescentando duas variáveis auxiliares ao arquivo **ViagraCovariate.sav** que usamos neste capítulo (veja a Seção 8.2.2 para ajuda com a codificação da variável auxiliar). Se você não souber o que fazer, incluí um conjunto completo de dados chamado **ViagraCovariate Dummy.sav** no *site* www.artmed.com.br. Para conduzir a análise, usamos o procedimento de regressão (veja o Capítulo 5) com a **libido** como saída (variável dependente na terminologia do SPSS), depois entramos com a libido do parceiro como um previsor no primeiro bloco (**partner**) e, então, entramos com as duas variáveis *dummies* (auxiliares) no segundo bloco. Em todos os casos, o método de entrada deve ser o **Enter** (consulte a Seção 5.7 para rever como entrar com variáveis em uma regressão hierárquica, mas lembre-se de usar o método **Enter** para ambos os blocos). O resumo do modelo de regressão resultante (Saída 9.7 do SPSS) mostra a aderência do modelo primeiro quando somente a covariável é usada e em seguida quando tanto a covariável quanto as variáveis auxiliares são utilizadas.

Material com direitos autorais

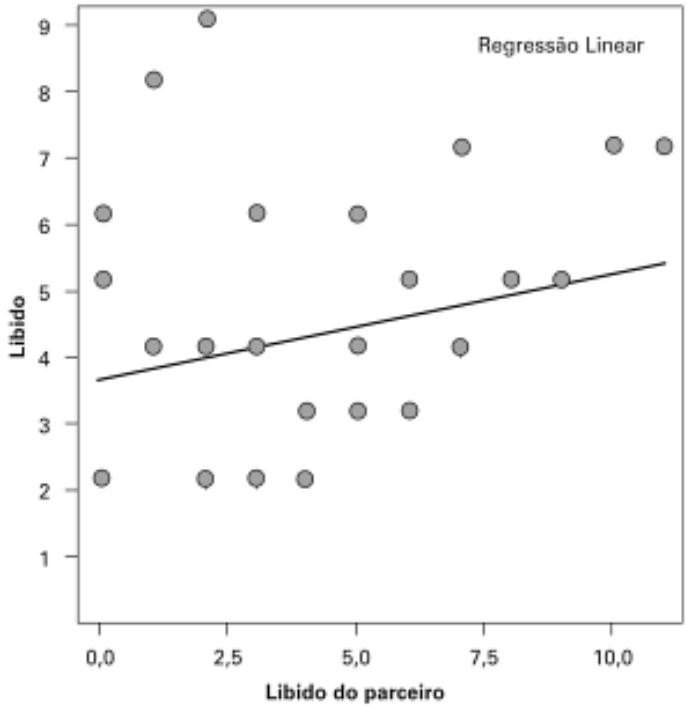


Figura 9.5 Diagrama de dispersão da libido do participante contra a libido do parceiro.

Portanto, a diferença entre os valores de R^2 (0,313 – 0,058 = 0,255) representa a contribuição individual da dose de Viagra. Assim, podemos dizer que a dose de Viagra é responsável por 25,5% da variação na libido enquanto a libido do parceiro é responsável por somente 5,8%. Essa informação adicional fornece uma percepção sobre a importância do Viagra. A próxima tabela é a da ANOVA, que é novamente dividida em duas seções. A metade superior representa apenas o efeito da covariável e a metade inferior representa todo o modelo (isto é, incluindo a covariável e as dosagens de Viagra). Note que na parte inferior da tabela da ANOVA (a parte do Modelo 2) todo o modelo (a libido do parceiro e as variáveis auxiliares) são responsáveis por 34,57 unidades da variância (SS_M), existem 110,97 unidades no total (SS_T) e a variância não-explicada (SS_R) é 76,22. A metade inferior, portanto, contém os mesmos valores que a tabela resumo da ANCOVA da saída 9.2 do SPSS. Esses valores são os mesmos da linha rotulada linha *corrected model* (modelo corrigido) da tabela resumo da ANCOVA que encontramos quando executamos a análise como ANCOVA (veja a Saída 9.2 do SPSS).

A Saída 9.8 do SPSS mostra o restante da análise de regressão. Essa tabela de coeficientes de regressão é mais interessante. Novamente, essa tabela é dividida em duas e, assim, a parte inferior da tabela se parece com todo o modelo. Quando a dosagem do Viagra é considerada com a covariável, o valor de *b* para a covariável é 0,483, o que corresponde ao valor estimado do parâmetro na ANCOVA (Saída 9.3 do SPSS). Os valores de *b* para as variáveis auxiliares representam as diferenças entre as médias dos grupos de baixa dosagem e o grupo-placebo (**dummy1 –** auxiliar 1) e o grupo de alta dosagem e o grupo-placebo (**dummy2 –** auxiliar 2) – veja a Seção 8.2.2 para saber por quê. As médias dos grupos de alta e baixa dosagem foram de 4,88 e 4,85, respectivamente, e a média do grupo-placebo

Quadro 9.2 *(Continuação)*

***Coefficients*[a] (Coeficientes)**

Model (Modelo)	*Unstandardized Coefficients* (Coeficientes não-padronizados)		*Standardized Coefficients* (Coeficientes padronizados)		
	B	*Std. Error* (Erro Padrão)	*Beta*	*t*	*Sig.*
1 *(Constant)* (Constante)	3.689	0.626		5.894	0.000
Partner's Libido (Libido do Parceiro)	0.158	0.120	0.240	1.311	0.201
2 *(Constant)* (Constante)	2.009	0.986		2.038	0.052
Partner's Libido (Libido do Parceiro)	0.483	0.196	0.737	2.472	0.020
Dummy Variable 1 (Placebo vs. Low & High) (Variável Auxiliar 1 (Placebo *versus* Baixa e Alta))	0.387	0.238	0.276	1.623	0.117
Dummy Variable 2 (Low vs. High) (Variável Auxiliar 2 (Baixa *versus* Alta Dosagem))	1.447	0.705	0.617	2.051	0.050

a. Variável dependente: Libido.

Execute a análise como descrito na Seção 9.5. A saída resultante irá iniciar com um resumo do modelo e uma tabela da ANOVA que deve ser idêntica àquelas na Saída 9.7 do SPSS (porque fizemos as mesmas coisas que anteriormente, a única diferença é como a variância do modelo é subsequentemente separada com os contrastes). Os coeficientes de regressão para as variáveis auxiliares serão diferentes porque agora especificamos códigos distintos.

A primeira variável auxiliar compara o grupo-placebo com os grupos de alta e baixa dosagem. Assim, ela compara a média ajustada do grupo-placebo (3,313) com o valor médio das médias ajustadas para os grupos de baixa e alta dosagem (3,027 + 5,920)/2 = 4,474. O valor de *b* para a variável auxiliar deve, portanto, ser diferente entre esses valores: 4,474 – 3.313 = 1,16. Entretanto, também descobrimos, numa seção complexa e chata (8.2.10.2) que esse valor é dividido pelo número de grupos dentro do contraste (isto é, 3), assim, ele será 1.16/3 = 0,387 (como ele será na saída). A estatística *t* associada não é significativa, indicando que o grupo-placebo não foi significativamente diferente da média combinada dos grupos do Viagra.

A segunda variável auxiliar compara os grupos de alta e baixa dosagem, e assim o valor de *b* deve ser a diferença entre as médias ajustadas desses grupos: 5,920 – 3,027 = 2,89. Também descobrimos na Seção 8.2.10.2 que esse valor é dividido pelo número dos grupos dentro do contraste (isto é, 2) e assim será 2,89/2 = 1,447 (como na saída). A estatística *t* associada é significativa (sua significância é exatamente 0,05), indicando que o grupo de alta dosagem produziu uma libido significativamente mais alta do que o grupo de baixa dosagem após o controle do efeito da libido do parceiro.

Isso ilustra como você pode aplicar os princípios da Seção 8.2.10 para a ANCOVA. Embora o SPSS não forneça uma interface fácil para fazer os contrastes planejados, eles podem ser feitos se você usar os menus de regressão em vez dos da ANOVA!

9.6 SUPOSIÇÕES ADICIONAIS DA ANCOVA ③

9.6.1 Homogeneidade dos parâmetros da regressão ③

Quando uma ANCOVA é realizada, olhamos para o relacionamento total entre a saída (variável dependente) e a covariável: ajustamos uma linha de regressão para todo o conjunto de dados, ignorando o grupo que a pessoa pertence. Ajustando esse modelo nós, portanto, assumimos que esse relacionamento é verdadeiro para todos os grupos de participantes. Por exemplo, se existe um relacionamento positivo

Material com direitos autorais

entre a covariável e a saída em um grupo, assumimos que existe um relacionamento positivo em todos os outros grupos também. Se, entretanto, o relacionamento entre a saída (variável dependente) e a covariável difere entre os grupos, todo o modelo de regressão não é preciso (ele não representa todos os grupos). Essa suposição é muito importante e é chamada de hipótese da *homogeneidade dos parâmetros da regressão*. A melhor maneira de entender essa suposição é imaginar um diagrama de dispersão para cada condição experimental com a covariável num eixo e a saída em outro. Se você calculou e traçou a linha de regressão para cada um desses diagramas de dispersão, deve descobrir que as linhas de regressão são mais ou menos semelhantes (isto é, os valores de *b* em cada grupo devem ser iguais).

A Figura 9.5 mostra os diagramas de dispersão que exibem o relacionamento entre a libido do parceiro (a covariável) e a saída (a libido do participante) para cada uma das três condições experimentais. Em cada diagrama de dispersão um ponto representa os dados de um participante específico e as linhas são os parâmetros de regressão para o grupo particular (isto é, eles resumem o relacionamento entre a libido e a libido do parceiro mostrada pelos pontos). Deve ficar claro que existe um relacionamento positivo (as linhas dos parâ-

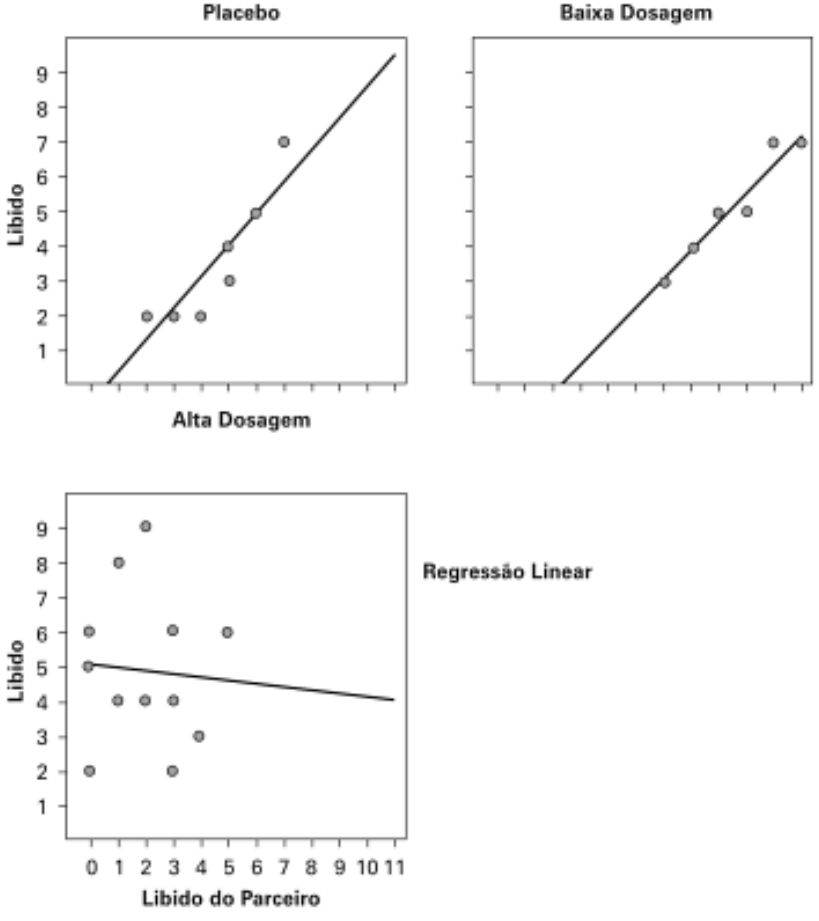


Figura 9.5 Diagrama de dispersão e linhas de regressão da libido do participante *versus* a libido do parceiro para cada uma das condições experimentais.

Material com direitos autorais

metros de regressão sobem da esquerda para a direita) entre a libido do parceiro e a libido do participante nas condições dos grupos-placebo e baixa dosagem. Entretanto, na condição de alta dosagem parece não haver relacão entre a libido dos participantes e a dos parceiros (os pontos estão aleatoriamente espalhados e, na verdade, a linha do parâmetro de regressão desce da esquerda para a direita indicando uma relação levemente negativa. Essa observação nos fornece um indício para duvidar que exista uma homogeneidade nos parâmetros de dispersão (porque o relacionamento entre a libido do participante e do seu parceiro não é consistente entre os três grupos-experimentais).

9.6.2 Testando a homogeneidade dos parâmetros da regressão no SPSS ③

Para testar a hipótese da homogeneidade dos parâmetros da regressão precisamos executar novamente a ANCOVA, mas dessa vez utilizando um modelo customizado. Acesse a caixa de diálogo principal como anteriormente e coloque as variáveis nos mesmos quadros que antes (a janela final deve ficar como a da Figura 9.1). Para customizar o modelo precisamos acessar a caixa de diálogo **model** (Figura 9.6) clicando em Model... (modelo).

Para customizar o modelo, clique no círculo rotulado de **Custom** (Customizar) a fim de ativar a caixa de diálogo (Figura 9.6). As variáveis especificadas na caixa de diálogo principal estão listadas no lado esquerdo e são seguidas por uma letra indicando o seu tipo (F = fator fixo, C = covariável). Para testar a homogeneidade dos parâmetros da regressão, precisamos especificar um modelo que inclui a interação entre a covariável e a variável dependente. Normalmente, a ANCOVA inclui somente o efeito principal da dosagem e parceiro e não inclui o termo de interação. Para testar o termo de interação é importante incluir os efeitos principais de dosagem e parceiros a fim de que o termo de interação seja testado controlando os efeitos principais.

Portanto, comece selecionando **dose** e **partner** (parceiro)(você pode selecionar ambos ao mesmo tempo). Onde diz **Build Term(s)** (Construir Termo(s)) existe um menu tipo lista de escolha. Clique em ▶ para acessar esse menu e, depois, clique em **Main effects** (Efeitos principais). Clique em ▶ a fim de mover os efeitos principais de **dose** e **partner** para o quadro denominado **Model** (Modelo). Depois precisamos especificar o termo de interação. Para tanto, selecione **dose** e **partner** ao mesmo tempo e clique em ▼ para acessar o menu tipo lista de escolha novamente, mas dessa vez

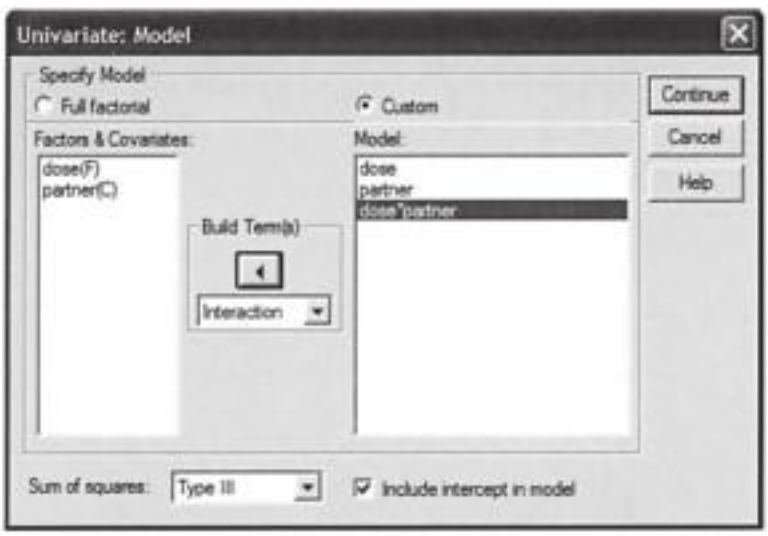


Figura 9.6 Caixa de diálogo *Model* (Modelo) do MLG univariado.

Material com direitos autorais

selecione **Interaction** (Interação). Clique em ▸ para mover a interação de **dose** e **partner** para a caixa rotulada de **Model** (Modelo). A caixa de diálogo final deve ficar igual a da Figura 9.6. Você pode descobrir mais sobre como especificar efeitos na Seção 10.3.2. Tendo especificado os dois efeitos principais e o termo de interação, clique em Continue (continuar) para retornar à caixa de diálogo principal e, então, em OK para executar a análise.

A Saída 9.9 do SPSS mostra a tabela de resumo principal para a ANCOVA incluindo o

Saída 9.9 do SPSS

***Tests of Between-Subjects Effects* (Testes de efeitos entre participantes)**

Dependent Variable: Libido (Variável Dependente: Libido)

Source (Fonte)	*Tipe III Sum of Squares* (Soma dos quadrados do tipo III)	*df* (gl)	*Mean Square* (Média ao quadrado)	*F*	*Sig.*
Corrected Model (Modelo Corrigido)	52.000[a]	5	10.400	4.233	0.007
Intercept (Intercepto)	1.706	1	1.706	0.694	0.413
DOSE (Dose)	40.068	2	20.034	8.154	0.002
Partner (Parceiro)	22.430	1	22.430	9.129	0.006
*Dose * Partner* (Dose * Parceiro)	17.250	2	8.625	3.510	0.046
Error (Erro)	58.967	24	2.457		
Total (Total)	683.000	30			
Corrected Total (Total corrigido)	110.967	29			

a. R quadrado = 0,469 (R quadrado ajustado = 0,358).

Dica da Samanta Ferrinho

- A Análise de Covariância (ANCOVA) compara várias médias, mas com controle do efeito de outras variáveis (uma ou mais), denominadas covariáveis; por exemplo, se você tem várias condições experimentais e quer controlar a idade dos participantes.
- Na tabela denominada *Tests of between-subjects effects* (Testes entre os efeitos dos participantes), olhe para a coluna denominada *Sig*. tanto para a covariável quanto para a variável independente (a variável de agrupamento); se o valor for menor do que 0,05, isso significa que a covariável tem relação significativa com a variável de saída (a variável dependente); para a variável de agrupamento, isso significa que as médias dos grupos são significativamente diferentes após o controle para o efeito que a covariável tem na saída.
- Como ocorreu com a ANOVA, se você gerou hipóteses específicas antes do experimento, utilize as comparações planejadas (*planned comparisons*); se você não tem hipóteses específicas, utilize os testes *post hoc*. Embora o SPSS deixe você especificar certos contrastes padrão, outras comparações planejadas deverão ser feitas com o recurso do procedimento da regressão no SPSS.
- Para contrastes e testes *post hoc*, consulte a coluna denominada *Sig*. para descobrir se suas comparações são significativas (isso irá ocorrer se o valor da significância for menor do que 0,05).
- Teste as mesmas hipóteses que foram feitas para a ANOVA, mas, além disso, teste a hipótese da *homogeneidade dos parâmetros da regressão*. Isso deve ser feito customizando o modelo da ANCOVA no SPSS.

Material com direitos autorais

termo de interação. Os efeitos da dosagem do Viagra e a libido do parceiro continuam significativos, mas estamos interessados mesmo é no termo de interação. Assim, olhe para o valor de significância da covariável da saída da interação (**dose*partner**); se esse efeito é significativo, a hipótese da homogeneidade dos parâmetros da regressão foi violada. O efeito aqui é significativo ($p < 0,05$); portanto, a hipótese não é sustentável. Embora esse achado não seja surpreendente devido ao padrão dos relacionamentos mostrados na Figura 9.5, ele gera certa preocupação sobre a análise principal, especialmente à luz dos achados contraditórios das comparações múltiplas. Esse exemplo ilustra porque é importante testar as hipóteses e não simplesmente aceitar os resultados de uma análise.

9.7 CALCULANDO O TAMANHO DE EFEITO ②

Vimos no Quadro 9.1 que podemos fazer o SPSS produzir o eta quadrado (η^2), o que é apenas o r^2 calculado do efeito entre grupos, SS_M, dividido pela variância total dos dados, SS_T. Essa medida do tamanho de efeito é levemente tendenciosa, por isso recomendo não usá-la. Como ocorre com a ANOVA, é recomendável utilizar o ômega quadrado (ω^2), uma versão menos tendenciosa do eta quadrado (veja a Seção 8.5). Entretanto, como vimos na Seção 8.5, ele somente pode ser calculado quando tivermos um mesmo número de participantes em cada grupo (e não temos isto aqui!). Assim, estamos um pouco confusos!

Mas nem tudo está perdido porque, como eu já disse muitas vezes, o tamanho de efeito global não é tão interessante quanto o tamanho de efeito para comparações mais específicas. Essas são fáceis de calcular porque selecionamos parâmetros de regressão (veja a Saída do SPSS 9.3) e, assim, temos a estatística *t* para a covariável e comparações entre os grupos de baixa e alta dosagem e dos grupos-placebo e de alta dosagem. Essas estatísticas *t* têm $N - 2$ graus de liberdade (veja o Capítulo 5), onde N é o tamanho total da amostra (nesse caso, 30). Podemos usar a mesma equação da Seção 7.5.5.[3]

$$r_{\text{contraste}} = \sqrt{\frac{t^2}{t^2 + gl}}$$

Portanto, obtemos (*t*s da Saída 9.3 do SPSS):

$$r_{\text{contraste}} = \sqrt{\frac{2,47^2}{2,47^2 + 28}} = \sqrt{\frac{6,10}{34,10}} = 0,42$$

$$r_{\text{Alta Dosagem vs. Placebo}} = \sqrt{\frac{-3,095^2}{-3,095^2 + 28}} = \sqrt{\frac{9,58}{37,57}} = 0,50$$

$$r_{\text{Alta Dosagem vs. Baixa}} = \sqrt{\frac{-2,051^2}{-2,051^2 + 28}} = \sqrt{\frac{4,21}{32,21}} = 0,36$$

Todos esses efeitos apresentam tamanhos de efeito de médio a grande (eles estão entre 0,3 e 0,5). Portanto, além de serem estatisticamente significativos, esses efeitos são descobertas importantes.

9.8 RELATANDO RESULTADOS ②

Relatar a ANCOVA é o mesmo que relatar uma ANOVA, exceto que precisamos relatar o efeito da covariável também. Para a covariável e o efeito experimental, fornecemos detalhes sobre a razão F e os graus de liberdade dos quais

[3] A rigor, devemos usar um procedimento um pouco mais elaborado quando os grupos não são iguais. Mas isto está fora do âmbito do livro. Rosnow, Rosenthal e Rubin (2000) fornecem um relato muito claro dessa situação.

Material com direitos autorais

10

ANOVA FATORIAL (MLG 3)

10.1 O QUE VOCÊ VAI APRENDER NESTE CAPÍTULO? ②

Nos dois capítulos anteriores, vimos situações nas quais tentamos testar diferenças entre grupos quando havia uma única variável independente (isto é, uma variável que havia sido manipulada). Entretanto, no início do Capítulo 8, afirmei que uma das vantagens da ANOVA era que poderíamos olhar para os efeitos de mais de uma variável independente (e como essas variáveis interagem). Este capítulo estende o que já sabemos sobre ANOVA à observação de situações onde existem duas variáveis independentes. Já vimos no capítulo anterior que é muito fácil incorporar uma segunda variável à estrutura da ANOVA quando aquela variável é uma variável contínua (isto é, não dividida em grupos), mas agora iremos ver situações onde existe uma segunda variável independente sistematicamente manipulada designando pessoas a diferentes situações.

10.2 TEORIA DA ANOVA FATORIAL (ENTRE GRUPOS) ②

10.2.1 Delineamentos fatoriais ②

A ANOVA **fatorial** é usada quando você tem duas ou mais variáveis independentes (essas variáveis são, às vezes, denominadas **fatores**, por isso o nome ANOVA *fatorial*). Existem vários tipos de delineamentos fatoriais:

- ***Delineamento fatorial independente:*** Quando existem diversas variáveis independentes ou previsoras e cada uma foi mensurada usando participantes diferentes (entre grupos).
- ***Delineamento fatorial relacionado:*** Quando diversas variáveis independentes ou previsoras foram mensuradas, mas os mesmos participantes foram usados em todas as condições (medidas repetidas).
- ***Delineamento misto***: Quando diversas variáveis independentes (ou previsoras) foram mensuradas; algumas foram mensuradas com diferentes participantes e outras usaram os mesmos participantes.

Esta seção estende a ANOVA de um fator para o caso fatorial e os capítulos subsequentes aos delineamentos de medidas repetidas, delineamentos fatoriais de medidas repetidas e finalmente, delineamentos misturados.

Material com direitos autorais

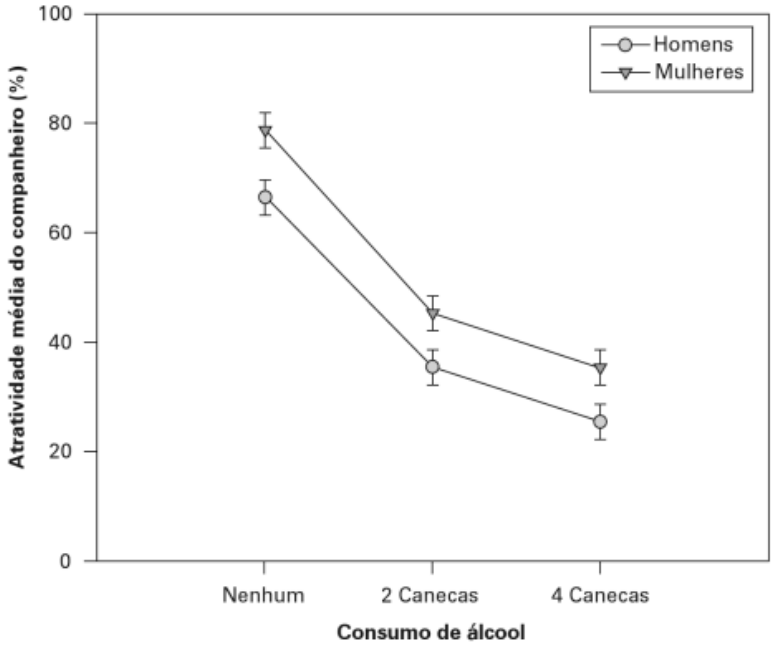


Figura 10.11 Um gráfico "sem" interação.

- Se as linhas num gráfico de interação se cruzam, obviamente elas não são paralelas e isso pode ser uma dica de que você tenha uma possível interação significativa. Entretanto, contrário à crença popular, nem *sempre* quando as linhas do gráfico de interação se cruzam, teremos uma interação significativa.

Outra complicação é que, algumas vezes, as pessoas desenham gráficos de barras em vez de gráficos de linhas. A Figura 10.12 mostra alguns gráficos de barras de interações entre duas variáveis independentes. Os painéis (a) e (b) exibem os dados do exemplo usados neste capítulo (na verdade, por que não traçá-los?). Como você pode ver, existem duas maneiras de apresentar os mesmos dados: o painel (a) mostra os dados quando os níveis de álcool estão localizados ao longo do eixo *x* e barras com cores diferentes são usadas para mostrar as médias para homens e mulheres e o painel (b) mostra um cenário oposto onde o gênero é traçado no eixo *x* e cores diferentes distinguem a dose do álcool. Ambos os gráficos mostram um efeito de interação. O que você está procurando são as diferenças entre as barras coloridas em pontos distintos do eixo *x*. Assim, para o painel (a) você observa as diferenças entre as barras escuras e claras do grupo sem álcool e, então, analisa a de 2 canecas e pergunta: "A diferença entre as barras é distinta daquela do grupo sem álcool?". Nesse caso, as barras parecem as mesmas das de sem álcool como as de 2 canecas: porém, sem interação. Entretanto, seguimos e observamos as de 4 canecas e perguntamos novamente: "A diferença entre as barras é diferente do que foi visto em qualquer outra condição?". Nesse caso, a resposta é sim: para sem álcool e 2 canecas as barras eram quase da mesma altura, mas para os 4 canecas a barra mais clara é muito mais alta do que a barra escura. Isso mostra uma interação: o padrão das respostas muda nas 4 canecas. O painel (b) mostra a mesma coisa, mas traçado de uma maneira inversa. Novamente, você observa o padrão das respostas. Assim, observamos os homens e vemos que

Material com direitos autorais

Tabela 10.2 Esquema de codificação para a ANOVA fatorial

Gênero	Álcool	Auxiliar 1 (Gênero)	Auxiliar 2 (Álcool)	Interação	Média
Masculino	Nenhum	0	0	0	68,875
Masculino	4 *canecas*	0	1	0	35,625
Feminino	Nenhum	1	0	0	60,625
Feminino	4 *canecas*	1	1	1	57,500

Para determinar o que os valores b representam nesse modelo, podemos fazer o mesmo que fizemos para o teste t e para a ANOVA de um fator independente; isto é, veja o que acontece quando inserimos valores dos nossos previsores (gênero e álcool)! Para começar, vejamos o que acontece quando observamos os homens que não ingeriram álcool. Nesse caso, o valor de gênero é 0, o valor de álcool é 0 e o valor da interação também é 0. A saída que previmos (como na ANOVA de um fator independente) é a média desse grupo (66,875), assim, nosso modelo fica:

$$\begin{aligned}\text{Atratividade}_i &= b_0 + b_1\text{Gênero}_i + b_2\text{Álcool}_i \\ &\quad + b_3\text{Interação}_i \\ \overline{X}_{\text{Homens,Sem}} &= b_0 + (b_1 \times 0) + (b_2 \times 0) \\ &\quad + (b_3 \times 0) \\ b_0 &= \overline{X}_{\text{Homens, Sem}} \\ b_0 &= 66{,}875\end{aligned}$$

Assim, a constante b_0 no modelo representa a média do grupo para o qual todas as variáveis estão codificadas como 0. Desse modo, é o valor da média da categoria base (nesse caso, homens que não beberam álcool). Agora, vamos ver o que acontece quando observamos as mulheres que não beberam álcool. Nesse caso, a variável gênero é 1 e as variáveis álcool e interação permanecem 0. Lembre, também, que o b_0 é a média dos homens que não beberam álcool. A saída é a média para as mulheres que não beberam álcool; portanto, a equação fica:

$$\begin{aligned}\text{Atratividade}_i &= b_0 + b_1\text{Gênero}_i + b_2\text{Álcool}_i \\ &\quad + b_3\text{Interação}_i \\ \overline{X}_{\text{Mulheres, Sem}} &= b_0 + (b_1 \times 1) + (b_2 \times 0) \\ &\quad + (b_3 \times 0) \\ \overline{X}_{\text{Mulheres, Sem}} &= \overline{X}_{\text{Homens, Sem}} + b_1 \\ b_0 &= \overline{X}_{\text{Mulheres, Sem}} - \overline{X}_{\text{Homens, Sem}} \\ &= 60{,}625 - 66{,}875 \\ &= -6{,}25\end{aligned}$$

Assim, b_1 no modelo representa a diferença entre homens e mulheres (para aqueles) que não beberam álcool. De uma forma mais geral podemos dizer que é o efeito do gênero para a categoria base do álcool (a categoria base sendo aquela codificada com 0, nesse caso, sem álcool). Agora, vamos observar os homens que beberam 4 canecas de álcool. Nesse caso, a variável gênero é 0, a variável do álcool é 1 e a variável de interação permanece 0. Podemos, também, trocar o b_0 pela média dos homens que beberam álcool. A saída é a média para os homens que beberam 4 canecas; portanto, a equação fica:

$$\begin{aligned}\overline{X}_{\text{Homens, 4 canecas}} &= b_0 + (b_1 \times 0) + (b_2 \times 1) \\ &\quad + (b_3 \times 0) \\ \overline{X}_{\text{Homens, 4 canecas}} &= b_0 + b_2 \\ \overline{X}_{\text{Homens, 4 canecas}} &= \overline{X}_{\text{Homens, Sem}} + b_2 \\ b_2 &= \overline{X}_{\text{Homens, 4 canecas}} - \overline{X}_{\text{Homens, Sem}} \\ &= 35{,}625 - 66{,}875 \\ &= -31{,}25\end{aligned}$$

Assim, o b_2 no modelo representa a diferença entre sem álcool e 4 canecas para os homens. De uma forma mais geral, é o efeito do álcool na categoria base de gênero (isto é, a categoria de gênero que foi codificada com um 0, nesse caso, homens). Finalmente, podemos observar as mulheres que beberam 4 canecas. Nesse caso, a variável gênero é 1, a variável álcool é 1 e a variável interação é também 1. Podemos, também, substituir b_0, b_1 pelo o que agora sabemos que eles representam. A saída é a média para as mulheres que beberam 4 canecas. Portanto, a equação fica:

Material com direitos autorais

11.2.5 O que fazer se a hipótese de esfericidade for violada? ②

Se os dados violam a hipótese de esfericidade, existem várias correções que podem ser aplicadas para produzir uma razão F válida. O SPSS produz três correções baseadas nas estimativas da esfericidade defendidas por Greenhouse e Geisser (1959) e Huynh e Feldt (1976). Ambas estimativas utilizam um fator de correção aplicada aos graus de liberdade usados para avaliar a razão F observada. O cálculo dessas estimativas está além do escopo deste livro (leitores interessados devem consultar Girden, 1992); precisamos saber somente que as três estimativas diferem. A *correção de Greenhouse e Geiser* (usualmente denotada como $\hat{\varepsilon}$) varia entre $1/(k-1)$ (onde k é o número das condições das medidas repetidas) e 1. Quanto mais próximo $\hat{\varepsilon}$ estiver de 1,00, mais homogêneas serão as variâncias das diferenças e portanto os dados estarão mais próximos da esfericidade. Por exemplo, numa situação na qual existem cinco condições, o limite mais baixo de $\hat{\varepsilon}$ será de $1/(5-1)$ ou 0,25 (conhecido como a estimativa limite inferior da esfericidade).

Huynh e Feldt (1976) relataram que quando a estimativa de Greenhouse e Geisser é maior do que 0,75, muitas hipóteses nulas falsas não são rejeitadas (isto é, a correção é muito conservadora) e Collier, Baker, Mandeville e Hayes (1967) mostraram que isso também aconteceu com uma estimativa de esfericidade de 0,90. Huynh e Feldt, portanto, propuseram sua própria correção menos conservadora (usualmente denotada como $\tilde{\varepsilon}$). Entretanto, Maxwell e Delaney (1990) relatam que $\tilde{\varepsilon}$ superestima a esfericidade. Stevens (1992), portanto, recomenda calcular a média das duas e ajustar o gl para esse valor médio. Girden (1992) recomenda que quando a estimativa da esfericidade é maior do que 0,75, a *correção de Huynh-Feldt* deve ser usada, mas quando a estimativa da esfericidade for menor do que 0,75 ou nada é conhecido sobre a esfericidade, a correção Greenhouse e Geisser deve ser utilizada. Veremos como usar esses valores no devido tempo.

Uma última opção para dados que violam a esfericidade é usar a estatística teste multivariada (MANOVA – veja o Capítulo 14), porque ela não depende da hipótese da esfericidade (veja O'Brien e Kaiser, 1985). A MANOVA evita a hipótese de esfericidade (e todas as considerações correspondentes sobre as razões F apropriadas e correções) usando um termo erro específico para contrastes com 1 gl e, portanto, cada contraste é somente associado com seu termo erro específico (em vez dos termos erro combinados usados na ANOVA). A MANOVA é tratada em profundidade no Capítulo 14, mas o procedimento de medidas repetidas do SPSS produz automaticamente estatísticas teste multivariadas.

Existe uma troca no poder do teste entre as abordagens univariada e multivariada (embora alguns autores argumentem que isso pode ser resolvido com um domínio adequado das técnicas – O'Brien e Kaiser, 1985). Davidson (1972) comparou o poder das técnicas univariadas ajustadas com as do T^2 do Hotelling (um teste estatístico MANOVA) e descobriu que a técnica univariada era relativamente menos poderosa para detectar pequenas mudanças confiáveis entre condições altamente correlacionadas quando outras condições menos correlacionadas estavam também presentes. Mendoza e colaboradores (1974) conduziram um estudo Monte Carlo comparando as técnicas univariadas e multivariadas sob violações de simetria e normalidade composta e descobriram que "à medida que o grau de violação da simetria composta aumenta, o poder empírico para os testes multivariados também aumenta. Em contraste, o poder para os testes univariados geralmente diminui" (p.174). Maxwell e Delaney (1990) observaram que o teste univariado é relativamente mais poderoso do que o teste multivariado à medida que n diminui e propôs

Material com direitos autorais

Tabela 12.1 Dados de **LooksOrPersonality.sav** (Att = Atraente, Av = Mediano, Ug = Feio)

	Muito Carisma			Pouco Carisma			Chato		
Aparência	Att	Av	Ug	Att	Av	Ug	Att	Av	Ug
Homem	86	84	67	88	69	50	97	48	47
	91	83	53	83	74	48	86	50	46
	89	88	48	99	70	48	90	45	48
	89	69	58	86	77	40	87	47	53
	80	81	57	88	71	50	82	50	45
	80	84	51	96	63	42	92	48	43
	89	85	61	87	79	44	86	50	45
	100	94	56	86	71	54	84	54	47
	90	74	54	92	71	58	78	38	45
Mulher	89	86	63	80	73	49	91	48	39
	89	91	93	88	65	54	55	48	52
	84	90	85	95	70	60	50	44	45
	99	100	89	80	79	53	51	48	44
	86	89	83	86	74	58	52	48	47
	89	87	80	83	74	43	58	50	48
	80	81	79	86	59	47	51	47	40
	82	92	85	81	66	47	50	45	47
	97	69	87	95	72	51	45	48	46
	95	92	90	98	64	53	54	53	45
	95	93	96	79	66	46	52	39	47

dos, use o arquivo **LooksOrPersonality.sav**. Primeiro temos que definir nossas variáveis de medidas repetidas, assim, acesse a caixa de diálogo **define factors** (definir fatores) usando o menu **Analyze⇒General Linear Model⇒Repeated Measures…** (Analisar⇒Modelo Linear Geral⇒Medidas Repetidas…). Da mesma forma que com a ANOVA de dois fatores de medidas repetidas (veja o capítulo anterior), precisamos nomear as nossas variáveis de medidas repetidas e especificar quantos níveis cada uma tem. Nesse caso, existem dois fatores entre sujeitos: **looks** (aparência) (atraente, mediano ou feio) e **charisma** (carisma) (muito carisma, pouco carisma e chato). Na caixa de diálogo **define factors**, troque a palavra **factor 1** pela palavra **looks**. Quando você tiver dado a esse fator de medidas repetidas um nome, diga ao computador que essa variável tem três níveis digitando o número 3 dentro da caixa **Number of Levels** (Número de níveis). Clique em Add (adicionar) para incluir essa variável na lista de variáveis de medidas repetidas; ela irá aparecer no quadro em branco no final da caixa de diálogo como **looks** (3) (aparência). Agora, repita esse processo com a segunda variável independente Insira a palavra **charisma** no espaço rotulado **Within-Subject Factor Name** e então, devido aos 3 níveis dessa variável, insira o número 3 no espaço rotulado **Number of Levels**. Clique em Add (incluir) para incluir essa variável na lista de fatores, que aparecerá como **charisma (3)**. A caixa de diálogo completa é apresentada na Figura 12.1. Depois de entrar com ambos os fatores dentre sujeitos, clique em Define (Definir) para ir à caixa de diálogo principal.

A caixa de diálogo principal é a mesma de quando fizemos a ANOVA fatorial de medidas repetidas no capítulo anterior (veja a Figura 12.2). No alto da caixa **Within-Subjects Variables** (Variáveis entre participantes), o SPSS afirma que existem dois fatores: **looks**

Material com direitos autorais

baixas cargas no segundo, enquanto o padrão é inverso para as três últimas variáveis. A diferença está somente nos valores das ponderações, que são menores porque as correlações entre as variáveis estão, agora, sendo levadas em conta. Esses coeficientes dos fatores podem ser utilizados para substituir os valores *b* na equação (15.4):

$$\begin{aligned}\text{Sociabilidade} &= 0{,}343\ \text{Conversa1} \\ &\quad + 0{,}376\ \text{Habilidades Sociais} \\ &\quad + 0{,}362\ \text{Interesse} \\ &\quad + 0{,}000\ \text{Conversa2} \\ &\quad - 0{,}037\ \text{Egoísmo} \\ &\quad + 0{,}039\ \text{Mentiroso} \\ \text{Sociabilidade} &= (0{,}343 \times 4) + (0{,}376 \times 9) \\ &\quad + (0{,}362 \times 8) + (0{,}000 \times 6) \\ &\quad - (0{,}037 \times 8) + (0{,}039 \times 6) \\ &= 7{,}59 \qquad (15.5) \\ \text{Consideração} &= 0{,}006\ \text{Conversa1} \\ &\quad - 0{,}020\ \text{Habilidades Sociais} \\ &\quad + 0{,}020\ \text{Interesse} \\ &\quad + 0{,}473\ \text{Conversa2} \\ &\quad + 0{,}437\ \text{Egoísmo} \\ &\quad + 0{,}405\ \text{Mentiroso} \\ \text{Consideração} &= (0{,}006 \times 4) - (0{,}020 \times 9) \\ &\quad + (0{,}020 \times 8) + (0{,}020 \times 6) \\ &\quad + (0{,}473 \times 6) + (0{,}437 \times 8) \\ &\quad + (0{,}405 \times 6) \\ &= 8{,}768\end{aligned}$$

A equação (15.5) mostra como esses escores dos coeficientes são utilizados para produzir dois fatores para cada pessoa. Nesse caso, o participante apresenta os mesmos escores em cada variável que foram utilizados na equação (15.4). Os escores resultantes são mais parecidos do que quando as cargas dos fatores foram utilizadas como ponderações porque as diferentes variâncias entre as seis variáveis agora estão sendo controladas. O fato dos valores serem bem semelhantes reflete que essa pessoa não apenas apresenta um escore alto em relação às variáveis representando sociabilidade, mas ela apresenta também pouca consideração para com os outros (isto é, ela apresenta altos escores igualmente nos dois fatores). Essa técnica de produzir escores de fatores assegura que os resultados tenham uma média de 0 e uma variância igual ao quadrado do coeficiente de correlação múltipla entre os escores dos fatores estimados e os verdadeiros valores dos fatores. Contudo, o problema com o método de regressão é que os escores podem estar correlacionados não apenas com outros fatores além daquele em que eles foram baseados, mas também com outros escores de diferentes fatores ortogonais.

15.2.3.2 Outros métodos ②

Para superar os problemas associados à técnica de regressão, dois ajustes foram propostos: *o método Bartlett* e o **método Anderson-Rubin**. O SPSS pode produzir escores dos fatores baseados em qualquer um desses métodos. O método Bartlett produz escores imparciais e que se correlacionam apenas com seu próprio fator. A média e o desvio padrão dos escores são os mesmos do método de regressão. Entretanto, os escores dos fatores podem, ainda, se correlacionar entre si. O método final é o de Anderson-Rubin, uma modificação do de Bartlett que produz escores dos fatores padronizados e não-correlacionados (eles apresentam uma média de 0 e um desvio padrão de 1). Tabachnick e Fidell (2001) concluem que o método de Anderson-Rubin é melhor quando escores não-correlacionados são requeridos, mas o método de regressão é preferido em outras circunstâncias simplesmente porque é o de mais fácil compreensão. Embora não seja crucial que você entenda os cálculos por trás dos métodos, é importante que você saiba o que os escores dos fatores representam: a saber, um escore múltiplo para cada indivíduo em um fator particular.

15.2.3.3 Usos dos escores dos fatores ②

Existem vários usos para os escores dos fatores. Primeiro, se o propósito da análise de fatores é reduzir um grande conjunto de dados em um subconjunto menor de variáveis, os escores dos fatores nos darão um escore indivi-

Material com direitos autorais

técnicas diferem nas estimativas da comunalidade utilizada. De forma simplista, a análise dos fatores deriva de um modelo matemático em que os fatores são estimados, enquanto a análise dos componentes principais apenas decompõe os dados originais em um conjunto de variáveis lineares (veja Dunteman, 1989, Capítulo 8, para mais detalhes sobre as diferenças entre os dois procedimentos). Assim, somente a análise dos fatores pode estimar os fatores subjacentes e ela tem como base várias hipóteses para que essas estimativas sejam precisas. A análise dos componentes principais se preocupa somente em determinar que componentes lineares existem dentro dos dados e como uma variável particular pode contribuir com aquele componente. Em termos de teoria, este capítulo é dedicado mais à análise dos componentes principais do que propriamente à análise dos fatores. A razão é que a análise dos componentes principais é um procedimento totalmente psicométrico, ela é conceitualmente menos complexa do que a análise dos fatores e apresenta muitas similaridades com a análise discriminante (descrita no capítulo anterior).

Entretanto, devemos considerar se as técnicas fornecem soluções diferentes para o mesmo problema. Baseados em uma extensiva revisão da literatura, Guadagnoli e Velicer (1988) concluíram que a solução gerada pela análise dos componentes principais difere um pouco daquela derivada da técnica da análise dos fatores. Na realidade, existem algumas circunstâncias nas quais essa afirmação não é verdadeira. Stevens (1992) resume os exemplos e conclui que com 30 ou mais variáveis e comunalidades maiores do que 0,7 para todas as variáveis, diferentes soluções são improváveis; entretanto, com menos do que 20 variáveis e com baixas comunalidades (< 0,4) podem ocorrer diferenças.

O lado menos interessante desse argumento é eloquentemente descrito por Cliff (1987), que observou que defensores da análise de fatores "insistem que a análise de componentes é, no melhor caso, uma análise de fatores comuns com algum erro acrescido e, no pior caso, uma bagunça irreconhecível de coisas das quais nada pode ser determinado" (p. 349). Os ânimos se exaltam nesse assunto, alguns argumentando que quando a análise dos componentes principais for utilizada ela não deveria ser descrita como uma análise dos fatores e que não deve atribuir um significado substantivo aos componentes resultantes. Entretanto, para os não estatísticos, a diferença entre um componente principal e um fator pode ser difícil de conceitualizar (ambos são modelos lineares) e as diferenças surgem principalmente nos cálculos.[3]

15.3.4 A teoria por trás da Análise de Componentes Principais ③

A análise dos componentes principais trabalha de uma maneira muito semelhante à MANOVA e à análise discriminante (veja o capítulo anterior). Embora não seja necessário entender os princípios matemáticos em detalhes, leitores do capítulo anterior podem ter se beneficiado das comparações entre as duas técnicas. Para aqueles que não leram o capítulo anterior, sugiro uma rápida consulta antes de continuar!

Na MANOVA, várias matrizes SSCP calculadas continham informações sobre os relacionamentos entre variáveis dependentes. Mencionei anteriormente que essas matrizes SSCP poderiam ser facilmente convertidas em matrizes variância-covariância, que apresentam a mesma informação, mas em forma de média (isto é, levando em conta o número de observações). Comentei, também, que dividindo cada elemento pelo desvio padrão pertinente, as matrizes variância-covariância se tornam padronizadas. O resultado é uma matriz de correlação. Na análise dos fatores, geralmente

[3] Por esse motivo, usei os termos *componentes* e *fatores* da mesma forma ao longo do capítulo. Embora esse uso dos termos injurie alguns estatísticos (e psicólogos), aposto que essas pessoas não irão ler este livro! Reconheço as diferenças metodológicas, mas é mais fácil para os estudantes se eu enfatizar as semelhanças entre as técnicas e não as diferenças.

Material com direitos autorais

tratamos com matrizes de correlação (embora seja possível analisar a matriz de variância-covariância também) e que essa matriz apresenta a mesma informação que a matriz SSCP na MANOVA. A diferença é apenas que a matriz de correlação é uma versão média da SSCP que foi padronizada.

Na MANOVA, usamos várias matrizes SSCP que representam diferentes componentes da variação experimental (a variação do modelo e a variação residual). Na análise de componentes principais, a matriz de covariâncias (ou correlações) não pode ser desmembrada dessa maneira (porque todos os dados vêm do mesmo grupo de participantes). Na MANOVA, acabamos olhando para as combinações lineares (*variates*) ou componentes da matriz SSCP que representam a razão entre a variância do modelo e a variância do erro. Essas combinações lineares (*variates*) são dimensões lineares que separam os grupos testados, e vimos que as variáveis dependentes foram representadas por esses componentes subjacentes. Resumindo, verificamos se os grupos poderiam ser separados por uma combinação linear das variáveis dependentes. Essas combinações lineares foram encontradas calculando os autovetores da SSCP. O número de combinações lineares obtido foi menor do que p (o número de variáveis dependentes) ou $k - 1$ (onde k é o número dos grupos). Na análise de componentes, fazemos algo semelhante (estou simplificando um pouco as coisas, mas isso lhe dará uma ideia básica). Isto é, pegamos uma matriz de correlação e calculamos as combinações lineares. Não há grupos de observação, assim, o número de combinações lineares obtidas será sempre igual ao número de variáveis medidas (p). As combinações lineares são descritas, como na MANOVA pelos autovetores associados à matriz de correlação. Os elementos dos autovetores são os pesos de cada variável na combinação linear (veja a Equação 14.5). Esses valores são as cargas dos fatores descritas anteriormente. O maior autovalor associado com cada um dos autovetores fornece um indicador único da importância de cada combinação linear (ou componente). A ideia básica é retermos fatores com autovalores relativamente grandes e ignoramos aqueles com autovalores relativamente pequenos.

Resumindo, a análise de componentes trabalha de uma maneira similar à MANOVA. Começamos com uma matriz representando os relacionamentos entre variáveis. Os componentes lineares (também chamados de combinações lineares (*variates*) ou fatores) daquela matriz são, então, calculados pela determinação dos autovalores da matriz. Esses autovalores são usados para calcular os autovetores, os elementos que fornecem as cargas de uma variável ou um fator em particular (isto é, eles são os valores b na equação (15.1)). O autovalor é, também, uma medida da importância do autovetor com o qual ele está associado.

15.3.5 Extração de fatores: autovalores e o diagrama de inclinação ②

Nem todos os fatores são extraídos numa análise e existe um debate sobre os critérios usados para decidir se um fator é estatisticamente importante. Mencionei acima que os autovalores associados a uma combinação linear (*variate*) indicam a importância daquele fator. Portanto, parece lógico que devem ser retidos somente fatores com autovalores grandes. Como decidimos se um autovalor é suficientemente grande ou não para representar um fator significativo? Bem, uma técnica advogada por Cattell (1966b) é traçar um gráfico de cada autovalor (eixo-y) contra o fator com o qual ele está associado (eixo-x). Esse gráfico é conhecido como um *diagrama de declividade*. Mencionei anteriormente que é possível obter tantos fatores quanto as variáveis existentes, e que cada uma tem um autovalor associado. Representando num gráfico os autovalores, a importância relativa de cada fator se torna aparente. Tipicamente, existirão alguns fatores com autovalores bem

Material com direitos autorais

relevantes. Portanto, descartamos alguma informação. Os fatores que retemos não irão explicar toda a variância dos dados (porque descartamos alguma informação), assim, as comunalidades após a extração serão sempre menores do que 1. Os fatores retidos não correspondem perfeitamente às variáveis originais – eles apenas refletem a variância comum presente nos dados. Se as comunalidades representam a perda de informação, elas são estatísticas importantes. Quanto mais próximas as comunalidades estão de 1, melhor os fatores explicarão os dados originais. É lógico que quanto mais fatores retidos, maiores serão as comunalidades (porque menos informação é descartada); portanto, as comunalidades são bons índices para verificar se poucos fatores foram retidos. Na verdade, com a *análise dos fatores por mínimos quadrados generalizados* e a *análise dos fatores por máxima verossimilhança*, você consegue uma medida estatística da aderência da solução dos fatores (veja o próximo capítulo para mais testes de aderência). Isso basicamente mede a proporção da variância que a solução por fatores explica (assim, pode ser pensado como uma comparação de comunalidades antes e após a extração).

Um conselho final: a sua decisão de quantos fatores extrair também dependerá de por que você está fazendo a análise em primeiro lugar (por exemplo, se você está tentando superar problemas de multicolinearidade na regressão, será melhor extrair mais fatores do que menos).

15.3.6 Melhorando a interpretação: a rotação de fatores ③

Depois de extrair os fatores, é possível calcular em que grau as variáveis se adaptam a esses fatores (isto é, calcular a carga da variável em cada fator). Em geral, você descobrirá que a maioria das variáveis tem cargas altas no fator mais importante e cargas baixas nos demais fatores. Essa característica torna a interpretação difícil, assim, uma técnica chamada de *rotação* de fatores é usada para fazer a distinção entre fatores. Se um fator for um eixo classificatório ao longo do qual as variáveis podem ser traçadas, a rotação dos fatores gira esses eixos de modo que as variáveis estejam carregadas ao máximo somente em um fator. A Figura 15.3 demonstra como esse processo funciona usando um exemplo

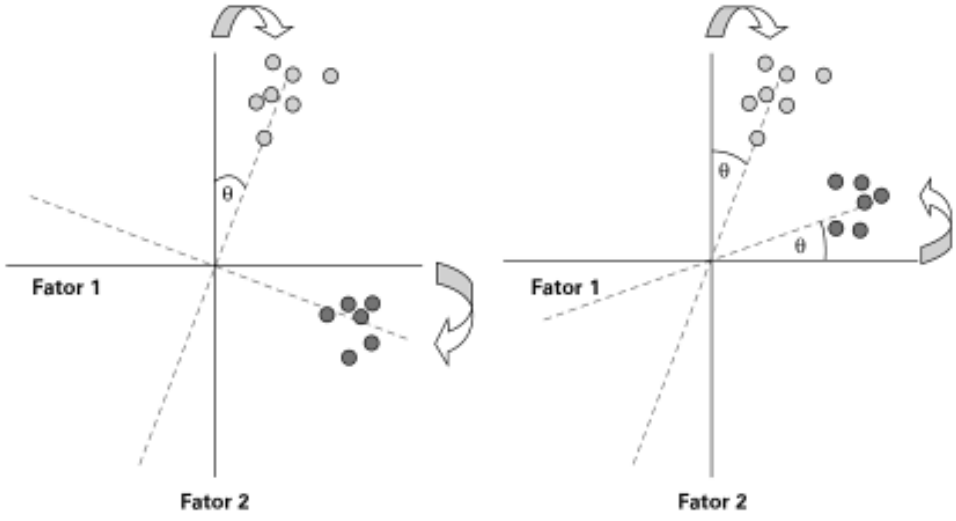


Figura 15.3 Representação esquemática da rotação de fatores. O gráfico da esquerda mostra a rotação ortogonal, enquanto o gráfico da direita mostra a rotação oblíqua (veja o texto para mais detalhes). θ é o ângulo pelo qual os eixos giram.

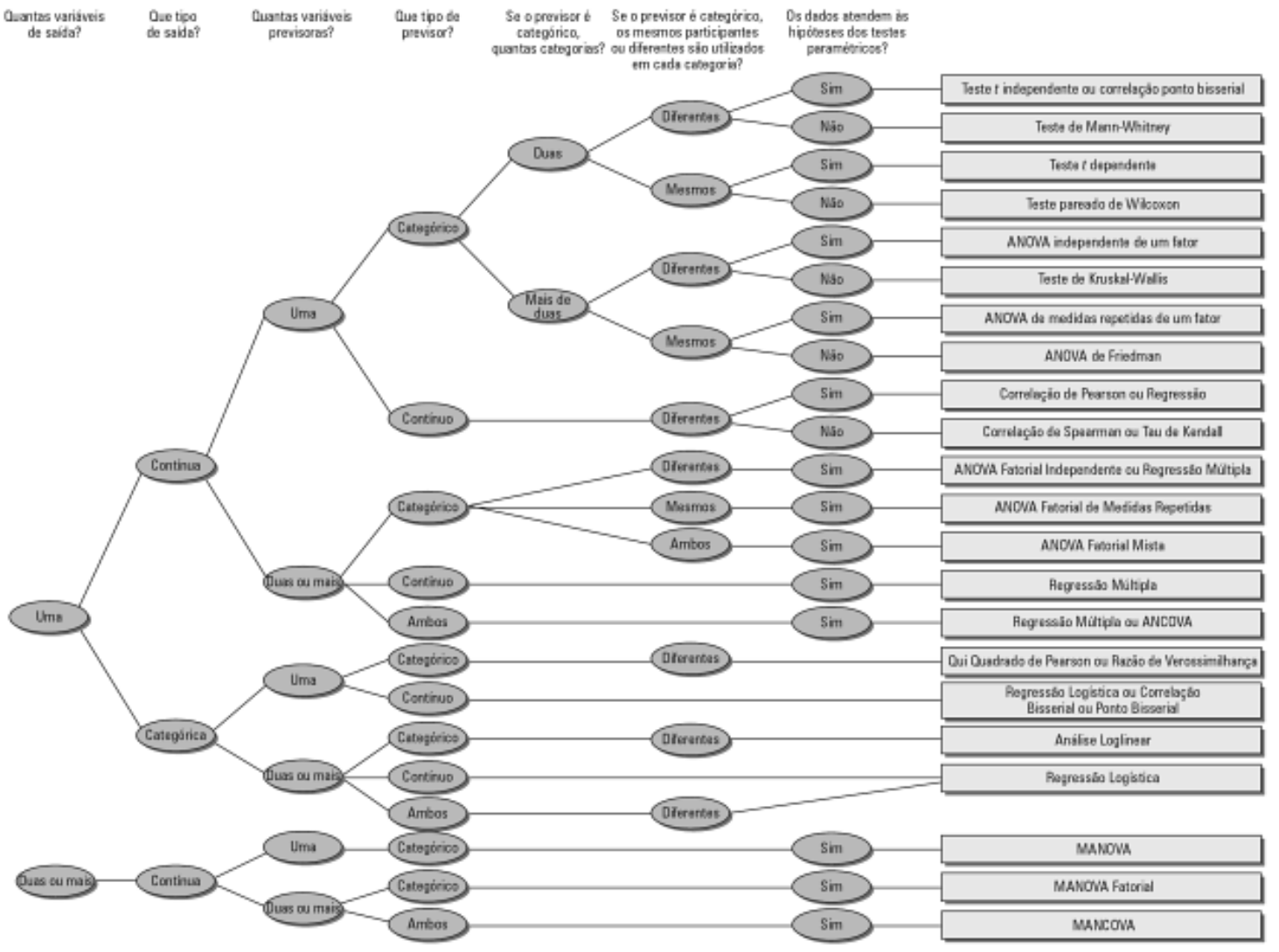
Material com direitos autorais

z	Larger Portion	Smaller Portion	y
0.38	0.64803	0.35197	0.3712
0.39	0.65173	0.34827	0.3697
0.40	0.65542	0.34458	0.3683
0.41	0.65910	0.34090	0.3668
0.42	0.66276	0.33724	0.3653
0.43	0.66640	0.33360	0.3637
0.44	0.67003	0.32997	0.3621
0.45	0.67364	0.32636	0.3605
0.46	0.67724	0.32276	0.3589
0.47	0.68082	0.31918	0.3572
0.48	0.68439	0.31561	0.3555
0.49	0.68793	0.31207	0.3538
0.50	0.69146	0.30854	0.3521
0.51	0.69497	0.30503	0.3503
0.52	0.69847	0.30153	0.3485
0.53	0.70194	0.29806	0.3467
0.54	0.70540	0.29460	0.3448
0.55	0.70884	0.29116	0.3429
0.56	0.71226	0.28774	0.3410
0.57	0.71566	0.28434	0.3391
0.58	0.71904	0.28096	0.3372
0.59	0.72240	0.27760	0.3352
0.60	0.72575	0.27425	0.3332
0.61	0.72907	0.27093	0.3312
0.62	0.73237	0.26763	0.3292
0.63	0.73565	0.26435	0.3271
0.64	0.73891	0.26109	0.3251
0.65	0.74215	0.25785	0.3230
0.66	0.74537	0.25463	0.3209
0.67	0.74857	0.25143	0.3187
0.68	0.75175	0.24825	0.3166
0.69	0.75490	0.24510	0.3144
0.70	0.75804	0.24196	0.3123
0.71	0.76115	0.23885	0.3101
0.72	0.76424	0.23576	0.3079
0.73	0.76730	0.23270	0.3056
0.74	0.77035	0.22965	0.3034
0.75	0.77337	0.22663	0.3011
0.76	0.77637	0.22363	0.2989
0.77	0.77935	0.22065	0.2966
0.78	0.78230	0.21770	0.2943
0.79	0.78524	0.21476	0.2920
0.80	0.78814	0.21186	0.2897
0.81	0.79103	0.20897	0.2874
0.82	0.79389	0.20611	0.2850
0.83	0.79673	0.20327	0.2827
0.84	0.79955	0.20045	0.2803
0.85	0.80234	0.19766	0.2780
0.86	0.80511	0.19489	0.2756
0.87	0.80785	0.19215	0.2732
0.88	0.81057	0.18943	0.2709
0.89	0.81327	0.18673	0.2685
0.90	0.81594	0.18406	0.2661
0.91	0.81859	0.18141	0.2637
0.92	0.82121	0.17879	0.2613
0.93	0.82381	0.17619	0.2589
0.94	0.82639	0.17361	0.2565
0.95	0.82894	0.17106	0.2541
0.96	0.83147	0.16853	0.2516
0.97	0.83398	0.16602	0.2492
0.98	0.83646	0.16354	0.2468
0.99	0.83891	0.16109	0.2444
1.00	0.84134	0.15866	0.2420
1.01	0.84375	0.15625	0.2396
1.02	0.84614	0.15386	0.2371
1.03	0.84849	0.15151	0.2347
1.04	0.85083	0.14917	0.2323
1.05	0.85314	0.14686	0.2299
1.06	0.85543	0.14457	0.2275
1.07	0.85769	0.14231	0.2251
1.08	0.85993	0.14007	0.2227
1.09	0.86214	0.13786	0.2203
1.10	0.86433	0.13567	0.2179
1.11	0.86650	0.13350	0.2155
1.12	0.86864	0.13136	0.2131
1.13	0.87076	0.12924	0.2107
1.14	0.87286	0.12714	0.2083
1.15	0.87493	0.12507	0.2059
1.16	0.87698	0.12302	0.2036
1.17	0.87900	0.12100	0.2012
1.18	0.88100	0.11900	0.1989
1.19	0.88298	0.11702	0.1965
1.20	0.88493	0.11507	0.1942
1.21	0.88686	0.11314	0.1919
1.22	0.88877	0.11123	0.1895
1.23	0.89065	0.10935	0.1872
1.24	0.89251	0.10749	0.1849
1.25	0.89435	0.10565	0.1826
1.26	0.89617	0.10383	0.1804
1.27	0.89796	0.10204	0.1781
1.28	0.89973	0.10027	0.1758
1.29	0.90147	0.09853	0.1736
1.30	0.90320	0.09680	0.1714
1.31	0.90490	0.09510	0.1691
1.32	0.90658	0.09342	0.1669
1.33	0.90824	0.09176	0.1647
1.34	0.90988	0.09012	0.1626
1.35	0.91149	0.08851	0.1604
1.36	0.91309	0.08691	0.1582
1.37	0.91466	0.08534	0.1561
1.38	0.91621	0.08379	0.1539
1.39	0.91774	0.08226	0.1518
1.40	0.91924	0.08076	0.1497
1.41	0.92073	0.07927	0.1476
1.42	0.92220	0.07780	0.1456
1.43	0.92364	0.07636	0.1435
1.44	0.92507	0.07493	0.1415
1.45	0.92647	0.07353	0.1394
1.46	0.92785	0.07215	0.1374
1.47	0.92922	0.07078	0.1354
1.48	0.93056	0.06944	0.1334
1.49	0.93189	0.06811	0.1315
1.50	0.93319	0.06681	0.1295
1.51	0.93448	0.06552	0.1276
1.52	0.93574	0.06426	0.1257
1.53	0.93699	0.06301	0.1238
1.54	0.93822	0.06178	0.1219
1.55	0.93943	0.06057	0.1200
1.56	0.94062	0.05938	0.1182
1.57	0.94179	0.05821	0.1163

(Continua)

z	Larger Portion	Smaller Portion	y
1.58	0.94295	0.05705	0.1145
1.59	0.94408	0.05592	0.1127
1.60	0.94520	0.05480	0.1109
1.61	0.94630	0.05370	0.1092
1.62	0.94738	0.05262	0.1074
1.63	0.94845	0.05155	0.1057
1.64	0.94950	0.05050	0.1040
1.65	0.95053	0.04947	0.1023
1.66	0.95154	0.04846	0.1006
1.67	0.95254	0.04746	0.0989
1.68	0.95352	0.04648	0.0973
1.69	0.95449	0.04551	0.0957
1.70	0.95543	0.04457	0.0940
1.71	0.95637	0.04363	0.0925
1.72	0.95728	0.04272	0.0909
1.73	0.95818	0.04182	0.0893
1.74	0.95907	0.04093	0.0878
1.75	0.95994	0.04006	0.0863
1.76	0.96080	0.03920	0.0848
1.77	0.96164	0.03836	0.0833
1.78	0.96246	0.03754	0.0818
1.79	0.96327	0.03673	0.0804
1.80	0.96407	0.03593	0.0790
1.81	0.96485	0.03515	0.0775
1.82	0.96562	0.03438	0.0761
1.83	0.96638	0.03362	0.0748
1.84	0.96712	0.03288	0.0734
1.85	0.96784	0.03216	0.0721
1.86	0.96856	0.03144	0.0707
1.87	0.96926	0.03074	0.0694
1.88	0.96995	0.03005	0.0681
1.89	0.97062	0.02938	0.0669
1.90	0.97128	0.02872	0.0656
1.91	0.97193	0.02807	0.0644
1.92	0.97257	0.02743	0.0632
1.93	0.97320	0.02680	0.0620
1.94	0.97381	0.02619	0.0608
1.95	0.97441	0.02559	0.0596
1.96	0.97500	0.02500	0.0584
1.97	0.97558	0.02442	0.0573
1.98	0.97615	0.02385	0.0562
1.99	0.97670	0.02330	0.0551
2.00	0.97725	0.02275	0.0540
2.01	0.97778	0.02222	0.0529
2.02	0.97831	0.02169	0.0519
2.03	0.97882	0.02118	0.0508
2.04	0.97932	0.02068	0.0498
2.05	0.97982	0.02018	0.0488
2.06	0.98030	0.01970	0.0478
2.07	0.98077	0.01923	0.0468
2.08	0.98124	0.01876	0.0459
2.09	0.98169	0.01831	0.0449
2.10	0.98214	0.01786	0.0440
2.11	0.98257	0.01743	0.0431
2.12	0.98300	0.01700	0.0422
2.13	0.98341	0.01659	0.0413
2.14	0.98382	0.01618	0.0404
2.15	0.98422	0.01578	0.0396
2.16	0.98461	0.01539	0.0387
2.17	0.98500	0.01500	0.0379
2.18	0.98537	0.01463	0.0371
2.19	0.98574	0.01426	0.0363
2.20	0.98610	0.01390	0.0355
2.21	0.98645	0.01355	0.0347
2.22	0.98679	0.01321	0.0339
2.23	0.98713	0.01287	0.0332
2.24	0.98745	0.01255	0.0325
2.25	0.98778	0.01222	0.0317
2.26	0.98809	0.01191	0.0310
2.27	0.98840	0.01160	0.0303
2.28	0.98870	0.01130	0.0297
2.29	0.98899	0.01101	0.0290
2.30	0.98928	0.01072	0.0283
2.31	0.98956	0.01044	0.0277
2.32	0.98983	0.01017	0.0270
2.33	0.99010	0.00990	0.0264
2.34	0.99036	0.00964	0.0258
2.35	0.99061	0.00939	0.0252
2.36	0.99086	0.00914	0.0246
2.37	0.99111	0.00889	0.0241
2.38	0.99134	0.00866	0.0235
2.39	0.99158	0.00842	0.0229
2.40	0.99180	0.00820	0.0224
2.41	0.99202	0.00798	0.0219
2.42	0.99224	0.00776	0.0213
2.43	0.99245	0.00755	0.0208
2.44	0.99266	0.00734	0.0203
2.45	0.99286	0.00714	0.0198
2.46	0.99305	0.00695	0.0194
2.47	0.99324	0.00676	0.0189
2.48	0.99343	0.00657	0.0184
2.49	0.99361	0.00639	0.0180
2.50	0.99379	0.00621	0.0175
2.51	0.99396	0.00604	0.0171
2.52	0.99413	0.00587	0.0167
2.53	0.99430	0.00570	0.0163
2.54	0.99446	0.00554	0.0158
2.55	0.99461	0.00539	0.0154
2.56	0.99477	0.00523	0.0151
2.57	0.99492	0.00508	0.0147
2.58	0.99506	0.00494	0.0143
2.59	0.99520	0.00480	0.0139
2.60	0.99534	0.00466	0.0136
2.61	0.99547	0.00453	0.0132
2.62	0.99560	0.00440	0.0129
2.63	0.99573	0.00427	0.0126
2.64	0.99585	0.00415	0.0122
2.65	0.99598	0.00402	0.0119
2.66	0.99609	0.00391	0.0116
2.67	0.99621	0.00379	0.0113
2.68	0.99632	0.00368	0.0110
2.69	0.99643	0.00357	0.0107
2.70	0.99653	0.00347	0.0104
2.71	0.99664	0.00336	0.0101
2.72	0.99674	0.00326	0.0099
2.73	0.99683	0.00317	0.0096
2.74	0.99693	0.00307	0.0093
2.75	0.99702	0.00298	0.0091
2.76	0.99711	0.00289	0.0088
2.77	0.99720	0.00280	0.0086

(Continua)


Quantas variáveis de saída?
Que tipo de saída?
Quantas variáveis previsoras?
Que tipo de previsor?
Se o previsor é categórico, quantas categorias?
Se o previsor é categórico, os mesmos participantes ou diferentes são utilizados em cada categoria?
Os dados atendem às hipóteses dos testes paramétricos?
Uma
Contínua
Categórica
Duas ou mais
Contínua
Uma
Duas ou mais
Categórico
Contínuo
Ambos
Duas
Mais de duas
Diferentes
Mesmos
Ambos
Sim
Não
Teste t independente ou correlação ponto bisserial
Teste de Mann-Whitney
Teste t dependente
Teste pareado de Wilcoxon
ANOVA independente de um fator
Teste de Kruskal-Wallis
ANOVA de medidas repetidas de um fator
ANOVA de Friedman
Correlação de Pearson ou Regressão
Correlação de Spearman ou Tau de Kendall
ANOVA Fatorial Independente ou Regressão Múltipla
ANOVA Fatorial de Medidas Repetidas
ANOVA Fatorial Mista
Regressão Múltipla
Regressão Múltipla ou ANCOVA
Qui Quadrado de Pearson ou Razão de Verossimilhança
Regressão Logística ou Correlação Bisserial ou Ponto Bisserial
Análise Loglinear
Regressão Logística
MANOVA
MANOVA Fatorial
MANCOVA