6

Desenhos de pesquisa com grupos independentes

Visão geral

No Capítulo 2, apresentamos os quatro objetivos da pesquisa em psicologia: descrição, previsão, explicação e aplicação. Os psicólogos usam métodos observacionais para desenvolver descrições detalhadas do comportamento, muitas vezes em ambientes naturais. Os métodos de pesquisa com uso de levantamentos permitem que os psicólogos descrevam as atitudes e opiniões das pessoas. Os psicólogos conseguem fazer previsões sobre o comportamento e processos mentais quando descobrem medidas e observações que covariam (correlações). A descrição e a previsão são essenciais para o estudo científico do comportamento, mas não são suficientes para entender as suas causas. Os psicólogos também procuram uma explicação – o "porquê" do comportamento. Chegamos a uma explicação científica quando identificamos as causas de um fenômeno. Os Capítulos 6, 7 e 8 concentram-se no melhor método de pesquisa existente para identificar relações causais – *o método experimental*. Analisaremos como o método experimental é usado para testar teorias psicológicas, bem como responder a questões de importância prática.

Como já enfatizamos, a melhor abordagem geral de pesquisa é a *abordagem multimétodos*. Podemos confiar mais em nossas conclusões quando obtemos respostas comparáveis para uma pergunta de pesquisa usando métodos diferentes. Diz-se que nossas conclusões têm *validade convergente*. Cada método tem limitações diferentes, mas os métodos têm potencialidades complementares que superam essas limitações. O principal ponto forte do método experimental é que ele é especialmente efetivo para estabelecer relações de causa e efeito. Neste capítulo, discutimos as razões por que os pesquisadores fazem experimentos e analisamos a lógica subjacente da pesquisa experimental. Nosso foco é em um desenho experimental bastante utilizado – o desenho de grupos aleatórios. Descrevemos os procedimentos para formar grupos aleatórios e as ameaças à interpretação que se aplicam especificamente ao desenho de grupos aleatórios. Depois, descrevemos os procedimentos que os pesquisadores usam para analisar e interpretar os resultados que obtêm em seus experimentos, e também investigamos como os pesquisadores estabelecem a validade externa dos resultados experimentais. Concluímos o capítulo com uma consideração sobre dois outros desenhos de pesquisa envolvendo grupos independentes: o desenho de grupos pareados e o desenho de grupos naturais.

Por que os psicólogos fazem experimentos

- Os pesquisadores fazem experimentos para testar hipóteses sobre as causas do comportamento.
- Os experimentos permitem que os pesquisadores decidam se um tratamento ou programa altera o comportamento efetivamente.

Uma das principais razões por que os psicólogos usam experimentos é para fazer testes empíricos das hipóteses que derivam de teorias psicológicas. Por exemplo, Pennebaker (1989) desenvolveu uma teoria que sustenta que manter para si pensamentos e sentimentos relacionados com experiências dolorosas pode ter um custo físico para o indivíduo. Segundo essa "teoria da inibição", é fisicamente estressante manter essas experiências para si mesmo.

Pennebaker e seus colegas fizeram muitos experimentos em que designavam um grupo de sujeitos para escrever sobre situações emocionais pessoais e outro grupo para escrever sobre tópicos superficiais. De maneira condizente com as hipóteses derivadas da teoria da inibição, os sujeitos que escreveram sobre tópicos emocionais tiveram melhores resultados em saúde do que os que escreveram sobre tópicos superficiais. Todavia, nem todos os resultados condiziam com a teoria da inibição. Por exemplo, estudantes que deviam dançar expressivamente com base em uma experiência emocional não tiveram os mesmos benefícios à saúde que estudantes que dançaram e escreveram sobre sua experiência. Pennebaker e Francis (1996) fizeram outro teste da teoria e demonstraram que as mudanças cognitivas que ocorrem ao se escrever sobre as experiências emocionais eram críticas para explicar os resultados positivos para a saúde.

Nossa descrição breve dos testes da teoria da inibição ilustra o processo geral envolvido quando os psicólogos fazem experimentos para testar uma hipótese derivada de uma teoria. Se os resultados do experimento condizem com o que a hipótese prevê, a teoria recebe amparo. Por outro lado, se os resultados diferem do que se esperava, a teoria talvez precise ser modificada, desenvolvendo-se e testando-se uma nova hipótese em outro experimento. Testar hipóteses e revisar teorias com base nos resultados de experimentos às vezes pode ser um processo longo e árduo, como combinar as peças de um quebra-cabeça para formar uma imagem completa. A inter-relação autoaperfeiçoadora entre os experimentos e as explicações propostas é uma ferramenta fundamental que os psicólogos usam para entender as causas das maneiras como nós pensamos, sentimos e agimos.

Os experimentos bem feitos também ajudam a resolver os problemas da sociedade, proporcionando informações vitais sobre a efetividade de tratamentos em uma ampla variedade de áreas. Esse papel dos experimentos tem uma longa história no campo da medicina (Thomas, 1992). Por exemplo, perto do começo do século XIX, a febre tifoide e o *delirium tremens* costumavam ser fatais. A prática médica padrão naquela época era tratar essas duas condições com sangria, purga e outras "terapias" semelhantes. Em um experimento para testar a efetividade desses tratamentos, os pesquisadores designavam um grupo aleatoriamente para receber o tratamento padrão (sangria, purga, etc.) e um segundo grupo que não recebia nada, apenas repouso na cama, boa nutrição e observação. Thomas (1992) descreve os resultados desse experimento como "inequívocos e estarrecedores" (p. 9): o grupo que recebeu o tratamento da época ficou pior do que o grupo que não foi tratado. Tratar essas condições usando as práticas do começo do século XIX era pior do que não tratá-las de modo al-

gum! Experimentos como esse contribuíram para o entendimento de que muitas condições médicas são autocontidas: a doença segue seu curso, e o paciente se recupera por conta própria.

A lógica da pesquisa experimental

- Os pesquisadores manipulam uma variável independente em um experimento para observar o efeito sobre o comportamento, conforme determinado pela variável dependente.
- O controle experimental permite que os pesquisadores façam a inferência causal de que a variável independente *causou* as mudanças observadas na variável dependente.
- O controle é o ingrediente essencial dos experimentos; o controle experimental é obtido por manipulação, mantendo as condições constantes e balanceando.
- Um experimento tem validade interna quando satisfaz as três condições necessárias para a inferência causal: covariação, relação de ordem temporal e eliminação de causas alternativas plausíveis.
- Quando ocorre confusão, existe uma explicação alternativa plausível para a covariação observada e, portanto, o experimento carece de validade interna. Explicações alternativas plausíveis são descartadas mantendo as condições constantes e balanceando.

Um experimento verdadeiro envolve a *manipulação* de um ou mais fatores e a *medição* (observação) dos efeitos dessa manipulação sobre o comportamento. Como vimos no Capítulo 2, os fatores que o pesquisador controla ou manipula são chamados de *variáveis independentes*. Uma variável independente deve ter pelo menos dois níveis (também chamados de condições). Um nível pode ser considerado a condição de "tratamento", e o segundo nível, condição de controle (ou comparação). As medidas usadas para observar o efeito (se houver) das variáveis independentes são chamadas de *variáveis dependentes*. Um modo de lembrar a distinção entre esses dois tipos de variáveis é entender que o resultado (a variável dependente) *depende* da variável independente.

Os experimentos são efetivos para testar hipóteses porque nos permitem exercer um grau relativamente elevado de controle em uma situação. Os pesquisadores usam controle em experimentos para que possam afirmar com confiança que a variável independente *causou* as mudanças observadas na variável dependente. As três condições necessárias para fazer uma inferência causal são covariação, relação de ordem temporal e eliminação de causas alternativas plausíveis (ver Capítulo 2).

A covariação ocorre quando observamos uma relação entre as variáveis independentes e dependentes de um experimento. A relação de ordem temporal se estabelece quando os pesquisadores manipulam uma variável independente e *depois* observam uma diferença subsequente no comportamento (i.e., a diferença no comportamento depende da manipulação). Finalmente, a eliminação de causas alternativas plausíveis ocorre por meio do uso de procedimentos de controle, principalmente por *manter as condições constantes e balancear*. Quando as três condições para uma inferência causal são satisfeitas, diz-se que o experimento tem **validade interna**, e podemos dizer que a variável independente *causou* a diferença de comportamento medida pela variável dependente.

Desenho de grupos aleatórios

- Em um desenho de grupos independentes, cada grupo de sujeitos participa de apenas uma condição da variável independente.
- A designação aleatória a condições é usada para formar grupos comparáveis, balanceando ou calculando a média das características dos sujeitos (diferenças individuais) entre as condições da manipulação da variável independente.

- Quando se usa designação aleatória para formar grupos independentes para os níveis da variável independente, o experimento é chamado de desenho de grupos aleatórios.

Em um **desenho de grupos independentes**, cada grupo de sujeitos participa de uma condição diferente da variável independente.[1] O desenho de grupos independentes mais efetivo é aquele que usa a **designação aleatória** de sujeitos a condições para formar grupos comparáveis antes de implementar a variável independente. Quando usamos designação aleatória às condições do estudo, chamamos o desenho de grupos independentes de **desenho de grupos aleatórios**. A lógica do desenho é clara. Os grupos são formados de modo a serem semelhantes em todas as características importantes no começo do experimento. A seguir, no experimento em si, os grupos são tratados igualmente, exceto no nível da variável independente. Assim, qualquer diferença entre os grupos em relação à variável dependente deve ser causada pela variável independente.

Exemplo de desenho de grupos aleatórios

A lógica do método experimental e da aplicação de técnicas de controle que produzem validade interna pode ser ilustrada em um experimento que investigou a insatisfação de garotas com seus corpos, realizado no Reino Unido por Dittmar, Halliwell e Ive (2006). Seu objetivo era determinar se a exposição a imagens de corpos muito magros fazia as garotas terem sentimentos negativos em relação a seus próprios corpos. Muitos experimentos realizados com sujeitos adolescentes e adultos demonstram que as mulheres relatam maior insatisfação consigo mesmas após a exposição a um modelo feminino magro, comparado com outros tipos de imagens. Dittmar e seus colegas tentaram determinar se efeitos semelhantes são observados para garotas com apenas 5 anos de idade. A imagem corporal muito magra que testaram era a boneca Barbie. Estudos antropológicos que comparam as proporções corporais da Barbie com mulheres reais revelam que a boneca tem proporções corporais bastante irreais, ainda que tenha se tornado um ideal sociocultural de beleza feminina (ver Figura 6.1).

No experimento, leu-se, para pequenos grupos de garotas (5 anos e meio a 6 anos e meio de idade), uma história sobre "Mira", que comprava roupas e se preparava para ir a uma festa de aniversário. À medida que liam a história, as garotas olhavam livros ilustrados com seis cenas relacionadas com a história. Em uma condição do experimento, os livros ilustrados tinham imagens da boneca Barbie nas cenas da história (p.ex., comprando roupas de festa, arrumando-se para a festa). Em uma segunda condição, os livros ilustrados tinham cenas semelhantes, mas a figura apresentada era a boneca "Emme". A boneca Emme é uma linda boneca, com proporções corporais mais realistas, representando o tamanho 16 nos Estados Unidos (ver Figura 6.2). Finalmente, na terceira condição do experimento, os livros não mostravam a Barbie ou a Emme (ou nenhum corpo), mas apresentavam imagens neutras relacionadas com a história (p.ex., vitrines de lojas de roupas, balões coloridos). Essas três versões dos livros ilustrados (Barbie, Emme, neutra) representavam três níveis da variável independente que foi manipulada no experimento. Como diferentes grupos de garotas participaram de cada nível da variável independente, o experimento é descrito como um desenho de grupos independentes.

Manipulação Dittmar e colaboradores (2006) usaram a técnica de controle por *manipulação* para testar suas hipóteses sobre a insatisfação corporal das garotas. As três condições da variável independente permitiram que os pesquisadores fizessem comparações relevantes para as suas hipóteses. Se testassem apenas a condição da Barbie, seria impossível determinar se as imagens influenciavam a insatisfação corporal das

☑ **Figura 6.1** Nos Estados Unidos, 99% das garotas com 3 a 10 anos têm pelo menos uma Barbie, e a garota típica tem em média oito Barbies (Rogers, 1999).

garotas de algum modo. Assim, a condição da imagem neutra criou uma comparação – um modo de verificar se a insatisfação corporal das garotas diferia dependendo de se olhavam uma imagem ideal magra ou uma imagem neutra. A condição da boneca Emme acrescentou uma comparação importante. É possível que *qualquer* imagem de corpo pudesse influenciar as percepções das garotas sobre si mesmas. Dittmar e seus colegas testaram a hipótese de que apenas ideais de corpo magro, representados pela Barbie, causariam insatisfação com o corpo.

Quando terminaram de ler a história, as garotas devolveram os livros ilustrados e preencheram um questionário adequado para sua faixa etária. Embora Dittmar e seus colegas tenham usado várias medidas para avaliar a satisfação das garotas com seus corpos, iremos nos concentrar em apenas uma medida, a Child Figure Rating Scale. Essa escala tem duas colunas com sete desenhos de formas corporais femininas, variando de muito magra a muito acima do peso. Cada garota devia primeiro colorir a figura na primeira coluna que parecesse mais com o seu corpo atual (uma medida da forma corporal percebida). Depois, na segunda coluna, as garotas deviam colorir a figura que mostrasse a maneira como gostariam de parecer (a forma corporal ideal). Falou-se que podiam escolher qualquer uma das figuras e que podiam escolher a mesma figura em cada coluna. O escore de insatisfação com a forma corporal, a variável dependente, foi calculado contando-se o número de figuras

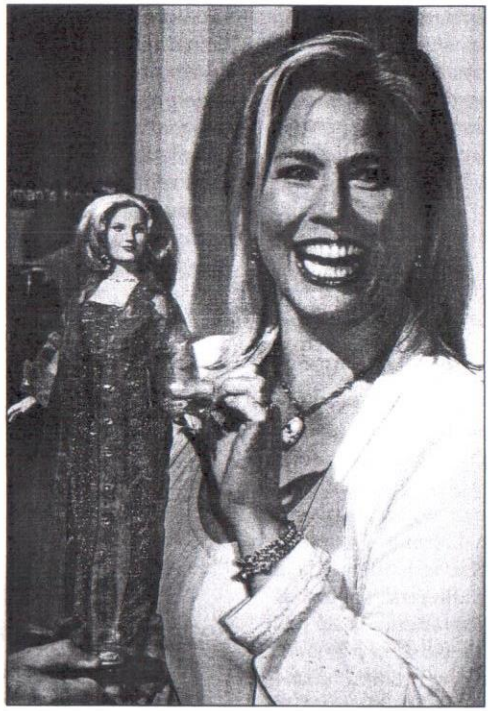

Figura 6.2 A boneca "Emme" foi lançada em 2002 para promover uma imagem corporal mais realista para as garotas. A boneca baseia-se em uma supermodelo americana chamada Emme.

entre a forma atual de cada garota e a sua forma ideal. Um escore de zero indicava que não havia insatisfação corporal, um escore negativo indicava o desejo de ser mais magra e um escore positivo indicava desejo de ter mais peso.

Os resultados desse experimento foram claros: as garotas expostas às imagens da Barbie ficaram mais insatisfeitas com sua forma corporal do que as garotas que foram expostas às imagens da Emme ou imagens neutras. O escore médio de insatisfação corporal para as 20 garotas na condição Emme e para as 20 garotas na condição de imagem neutra foi zero. Em comparação, o escore médio de insatisfação para as 17 garotas na condição da imagem da Barbie foi de -0,76, indicando seu desejo de serem mais magras. Por meio da técnica de controle da manipulação, as primeiras duas exigências para a inferência causal foram cumpridas no experimento: (1) diferenças na insatisfação corporal das garotas covariaram com as condições do experimento, e (2) a insatisfação corporal ocorreu após olharem as imagens (relação de ordem temporal). O terceiro requisito para a inferência causal, descartar explicações alternativas, foi cumprido no experimento mantendo-se as condições constantes e balanceando.

Condições constantes No experimento de Dittmar e colaboradores, vários fatores que poderiam ter afetado as atitudes das garotas para com seus corpos foram mantidos iguais nas três condições. Todas as garotas ouviram a mesma história sobre fazer

compras e a festa de aniversário, e olharam livros ilustrados pela mesma quantidade de tempo. Todas receberam as mesmas instruções no decorrer do experimento e receberam exatamente o mesmo questionário ao final. Os pesquisadores usam *condições constantes* para garantir que a variável independente seja o *único* fator que difira sistematicamente entre os grupos.

Se os três grupos tivessem diferido em um fator além dos livros ilustrados, teria sido impossível interpretar os resultados do experimento. Suponhamos que os participantes na condição da Barbie tivessem ouvido uma história diferente, por exemplo, uma história sobre uma Barbie magra e popular. Não saberíamos se a diferença observada na insatisfação corporal das garotas se deveria ao fato de verem as imagens da Barbie ou à história diferente. Quando se permite que a variável independente de interesse e uma variável diferente, potencialmente independente, covariem, existe uma *confusão*. Quando não existem variáveis confundidoras, o experimento tem *validade interna*.

Manter as condições constantes é uma técnica de controle que os pesquisadores usam para evitar confusões. Mantendo constante a história que as garotas ouviram nas três condições, Dittmar e seus colegas evitaram confusões com esse fator. De um modo geral, um fator que é mantido constante possivelmente covaria com a variável independente manipulada. Mais importante ainda, um fator que é mantido constante não muda, de modo que também não pode covariar com a variável dependente. Assim, os pesquisadores podem descartar fatores que são mantidos constantes como causas potenciais para os resultados observados.

Todavia, é importante reconhecer que controlamos apenas aqueles fatores que podem influenciar os comportamentos que estamos estudando – que consideramos como causas alternativas *plausíveis*. Por exemplo, Dittmar e colaboradores mantiveram constante a história que as garotas ouviram em cada condição. Todavia, é improvável que tenham controlado fatores como a temperatura da sala entre as condições, pois não seria provável que a temperatura afetasse a imagem corporal (pelo menos variando apenas alguns graus). Não obstante, devemos estar sempre alertas para a possibilidade de que pode haver fatores de confusão em nossos experimentos, cuja influência não tenhamos previsto ou considerado.

Balanceamento De forma clara, uma das chaves para a lógica do método experimental é formar grupos comparáveis (semelhantes) no começo do experimento. Os participantes de cada grupo devem ser comparáveis em termos de diversas características, como sua personalidade, inteligência, e assim por diante (também conhecidas como *diferenças individuais*). A técnica de controle por *balanceamento* é necessária porque esses fatores muitas vezes não podem ser mantidos constantes. O objetivo da designação aleatória é estabelecer grupos equivalentes de sujeitos, balanceando ou calculando a média das diferenças individuais entre as condições. O desenho de grupos aleatórios usado por Dittmar e colaboradores (2006) pode ser descrito da seguinte maneira:

Estágio 1	Estágio 2	Estágio 3
R_1	X_1	O_1
R_2	X_2	O_1
R_3	X_3	O_1

onde R_1, R_2 e R_3 referem-se à designação aleatória dos sujeitos às três condições independentes do experimento; X_1 é um nível de uma variável independente (p.ex., Barbie), X_2 é o segundo nível da variável independente (p.ex., Emme) e X_3 é o terceiro nível da variável independente (p.ex., imagens neutras). Faz-se uma observação do comportamento (O_1) em cada grupo.

No estudo de Dittmar e colaboradores (2006) sobre a imagem corporal das garotas, se as participantes que olharam as imagens da Barbie tivessem mais sobrepeso ou tivessem mais bonecas Barbies do que as que olharam imagens da Emme ou neutras,

haveria uma explicação alternativa plausível para os resultados. É possível que estar com sobrepeso ou ter mais Barbies, não a versão das imagens, explicasse por que as participantes na condição da Barbie sentiram mais insatisfação com seus corpos. (Na linguagem do pesquisador, haveria uma variável confundidora.) De maneira semelhante, diferenças individuais existentes na insatisfação corporal das garotas *antes* do experimento poderiam ser uma explicação alternativa razoável para os resultados do estudo. Todavia, usando designação aleatória para balancear essas diferenças individuais entre os grupos, podemos logicamente descartar a explicação alternativa de que as diferenças que obtivemos entre os grupos em relação à variável dependente se devem a características dos participantes.

Quando balanceamos um fator como o peso corporal, tornamos os três grupos equivalentes em termos de seu peso corporal *médio*. Observe que isso difere de manter o peso corporal constante, que exigiria que todas as garotas do estudo tivessem o mesmo peso. De maneira semelhante, balancear o número de bonecas Barbie das garotas nos três grupos significaria que o número *médio* de bonecas nos três grupos é o mesmo, e não que a quantidade de bonecas que cada garota possui é mantida em um dado número constante. A vantagem da designação aleatória é que *todas* as diferenças individuais são balanceadas, e não apenas aquelas que mencionamos. Portanto, podemos descartar explicações alternativas devidas a *qualquer* diferença individual entre as participantes.

Em suma, Dittmar e seus colegas concluíram que a exposição a imagens corporais magras, como a Barbie, *torna* as garotas insatisfeitas com seus próprios corpos. Eles conseguiram chegar a essa conclusão porque

- manipularam uma variável independente que variava as imagens que as garotas olhavam,
- descartaram outras explicações plausíveis mantendo as condições relevantes constantes e
- balancearam diferenças individuais entre os grupos por meio da designação aleatória às condições.

O Quadro 6.1 sintetiza como Dittmar e seus colegas aplicaram o método experimental, especificamente o desenho com grupos aleatórios, ao seu estudo sobre a imagem corporal de garotas.

Randomização em bloco

- A randomização em bloco equilibra as características dos sujeitos e variáveis confundidoras potenciais que ocorrem na implementação do experimento, e cria grupos de mesmo tamanho.

Um procedimento comum para a designação aleatória é a **randomização em bloco**. Vamos primeiro descrever exatamente como se faz randomização aleatória, e depois analisar o que ela faz. Suponhamos que temos um experimento com cinco condições (rotuladas, por conveniência, como A, B, C, D e E). Forma-se um "bloco" com uma ordem aleatória de todas as cinco condições:

Um bloco de condições	→	Ordem aleatória de condições
A B C D E		C A E B D

Na randomização em bloco, designamos os sujeitos a condições um bloco de cada vez. Em nosso exemplo com cinco condições, cinco sujeitos devem completar o primeiro bloco, com um sujeito em cada condição. Os próximos cinco sujeitos seriam designados a uma das cinco condições para completar um segundo bloco, e assim por diante. Se quiséssemos ter 10 sujeitos em cada uma das cinco condições, haveria 10 blocos no protocolo de blocos randomizados, cada um consistindo de um arranjo aleatório das cinco condições. O procedimento é ilustrado a seguir para os primeiros 11 participantes.

☑ Quadro 6.1

Síntese do experimento sobre a imagem corporal de garotas

Síntese do procedimento experimental. Garotas pequenas (5 anos e meio a 6 anos e meio de idade) foram designadas para olhar três livros ilustrados diferentes, enquanto escutavam uma história. Depois de olharem os livros, as participantes responderam perguntas sobre sua imagem corporal.

Variável independente. Versão do livro ilustrado observada pelas participantes (imagens da Barbie, Emme e neutras).

Variável dependente. Insatisfação corporal, medida avaliando-se a diferença entre a imagem corporal das garotas e sua imagem corporal ideal.

Explicação de procedimentos de controle

Manter condições constantes. As garotas nas três condições ouviram a mesma história, receberam as mesmas instruções e responderam as mesmas perguntas no final.

Balanceamento. As diferenças individuais entre as garotas foram balanceadas pela designação aleatória a diferentes condições experimentais.

Explicação da lógica experimental proporcionando evidências para a causalidade

Covariação. Observou-se que a insatisfação corporal das garotas variou com a condição experimental.

Relação de ordem temporal. A versão do livro ilustrado foi manipulada antes de se aferir a insatisfação corporal.

Eliminação de causas alternativas plausíveis. Os procedimentos de controle de manter as condições constantes e balancear diferenças individuais pela designação aleatória protegeram contra possíveis fatores de confusão.

Conclusão. A exposição a imagens corporais muito magras (os livros ilustrados com a boneca Barbie) causou insatisfação corporal.

(Baseado em Dittmar, Halliwell e Ive, 2006).

☑ Exercício I

Neste exercício, você deve responder as perguntas após esta breve descrição de um experimento.

Bushman (2005) investigou se a memória das pessoas para a publicidade é afetada pelo tipo de programa de televisão a que assistem. Os participantes (N = 336, idades 18-54) foram designados aleatoriamente para assistir a quatro tipos de programas de televisão: violento (p.ex., *Cops*), sexualidade explícita (p.ex., *Sex and the City*), violência e sexo (p.ex., *CSI Miami*) ou neutro (p.ex., *America's Funniest Animals*). Dentro de cada programa, foram embutidos os mesmos 12 comerciais (30 segundos). Para garantir que os participantes tivessem a mesma exposição às marcas representadas nos comerciais, os pesquisadores selecionaram marcas relativamente desconhecidas (p.ex., "Dermoplast", "José Olé"). Três intervalos comerciais, cada um com quatro anúncios, foram colocados aproximadamente aos 12, 24 e 36 minutos de cada programa, sendo usadas duas ordens aleatórias para os anúncios. Os sujeitos foram testados em grupos pequenos, e cada sessão foi realizada em um local confortável, onde os sujeitos sentavam em cadeiras estofadas e recebiam refrigerantes e petiscos. Depois de assistirem ao programa, os participantes receberam testes de memória de surpresa para o conteúdo dos comerciais. Os resultados indicaram que a memória para as marcas anunciadas foi pior quando o programa de televisão continha violência ou sexo. O comprometimento da memória para a publicidade foi maior para programas que continham material sexualmente explícito.

1. Que aspecto do experimento Bushman controlou usando manipulação?
2. Que aspecto do experimento Bushman controlou mantendo as condições constantes?
3. Que aspecto do experimento Bushman controlou usando balanceamento?

Bushman, B. J. (2005). Violence and sex in television programs do not sell products in advertisement. *Psychological Science, 16*, 702-708.

Blocos	Participantes		Condição	
A B D	1) Cara	→	C	Primeiro bloco
C D A B	2) Andy	→	A	
B E A C	3) Jacob	→	E	
B A C E D	4) Molly	→	B	
A C E D B	5) Emily	→	D	
A D E B C	6) Eric	→	E	Segundo bloco
B C A D E	7) Anna	→	C	
B C A E B	8) Laura	→	D	
D B C A	9) Sarah	→	A	
E B D A	10) Lisa	→	B	
	11) Tom	→	D	

E assim por diante para 50 participantes

Existem várias vantagens quando se usa a randomização em bloco para designar sujeitos aleatoriamente a grupos. Primeiramente, a randomização em bloco produz grupos de mesmo tamanho. Isso é importante porque o número de observações em cada grupo afeta a fidedignidade da análise descritiva para cada grupo, e é desejável que a fidedignidade dessas medidas seja comparável entre os grupos. A randomização em bloco faz isso. Em segundo lugar, a randomização em bloco controla as variáveis relacionadas com o tempo. Como os experimentos podem levar uma quantidade substancial de tempo para serem concluídos, alguns participantes podem ser afetados por algo que ocorra no decorrer do período de implementação do experimento. Na randomização em bloco, cada condição é testada em cada bloco, de modo que essas variáveis ligadas ao tempo são balanceadas entre as condições do experimento. Se, por exemplo, ocorre um fato traumático em um *campus* universitário onde se está conduzindo um experimento, o número de sujeitos que passaram pela experiência será equivalente em cada condição, sendo usada randomização em bloco. Pressupomos, então, que os efeitos da situação sobre o comportamento dos participantes sejam equivalentes, em média, entre as condições. A randomização em bloco também atua de maneira a balancear outras variáveis relacionadas com o tempo, como mudanças nos indivíduos que conduzem o experimento ou até mudanças nas populações de onde são tirados os sujeitos. Por exemplo, usando um protocolo com randomização em bloco, pode-se fazer um experimento perfeitamente aceitável com estudantes dos semestres de outono e primavera. A vantagem da randomização em bloco é que ela equilibra (ou usa a média) qualquer característica dos participantes (incluindo os efeitos de fatores relacionados com o tempo) entre as condições do experimento.

Se você quiser praticar o procedimento de randomização em bloco, responda o Desafio 1A ao final do capítulo.

Ameaças à validade interna

- Designar grupos inteiros aleatoriamente a diferentes condições da variável independente cria uma confusão potencial, em decorrência de diferenças preexistentes entre os participantes dos grupos inteiros.
- A randomização em bloco aumenta a validade interna, balanceando variáveis externas entre as condições da variável independente.
- A perda seletiva de sujeitos, mas não a perda mecânica de sujeitos, ameaça a validade interna de um experimento.
- Grupos controle com placebo são usados para controlar o problema das características de demanda, e os experimentos duplos-cegos controlam as características de demanda e os efeitos do experimentador.

Vimos que a *validade interna* é o grau em que diferenças no comportamento em relação a uma variável dependente podem ser atribuídas clara e definitivamente ao efeito de uma variável independente, em vez de alguma outra variável não controlada. Essas variáveis não controladas costumam ser citadas como **ameaças à validade interna**. Elas são explicações alternativas potenciais para os resultados de um estudo. Para fazer uma inferência clara de causa e efeito sobre uma variável independente, as ameaças à valida-

de interna devem ser controladas. A seguir, descreveremos diversos problemas na pesquisa experimental que podem resultar em ameaças à validade interna, bem como métodos para controlar essas ameaças.

Testando grupos inteiros A designação aleatória é usada para formar grupos comparáveis no desenho de grupos aleatórios. Todavia, existem ocasiões em que são formados grupos *in*comparáveis, mesmo que pareça ter sido usada designação aleatória. Esse problema ocorre quando grupos inteiros (e não indivíduos) são designados aleatoriamente às condições do experimento. Os grupos inteiros são formados antes do começo do experimento. Por exemplo, turmas diferentes de uma disciplina de introdução à psicologia são grupos inteiros. Os estudantes não são designados de forma aleatória a diferentes turmas de introdução à psicologia (embora às vezes o horário das classes pareça aleatório!). Os estudantes muitas vezes decidem estar em uma determinada turma por causa do horário das aulas, do professor, de amigos que estarão naquela aula, e de vários outros fatores. Se um pesquisador designasse diferentes turmas aleatoriamente a níveis de uma variável independente, poderia haver confusão, pela testagem de grupos inteiros.

A fonte da confusão devida ao uso de grupos incomparáveis ocorre quando os indivíduos diferem sistematicamente entre os grupos inteiros. Por exemplo, estudantes que decidem cursar introdução à psicologia na turma das 8 da manhã podem ser diferentes dos que preferem a turma das 14 horas. A designação aleatória desses grupos inteiros às condições experimentais simplesmente não seria suficiente para balancear as diferenças sistemáticas entre os grupos inteiros. Essas diferenças sistemáticas entre os dois grupos inteiros quase sempre ameaça a validade interna do experimento. A solução para esse problema é simples – não usar grupos inteiros em um desenho de grupos aleatórios.

Balanceando variáveis externas Diversos fatores em um experimento podem variar como consequência de considerações práticas ao executar o estudo. Por exemplo, para implementar um experimento mais rapidamente, o pesquisador deve usar vários experimentadores diferentes para testar pequenos grupos de participantes. Os tamanhos dos grupos e os próprios experimentadores se tornam variáveis potencialmente relevantes que poderiam confundir o experimento. Por exemplo, se todos os indivíduos no grupo experimental fossem testados por um experimentador e todos os do grupo controle fossem testados por outro, os níveis da variável independente pretendida seriam confundidos com os dois experimentadores. Não conseguiríamos determinar se uma diferença observada entre os dois grupos se devia à variável independente ao ou fato de que experimentadores diferentes testaram os sujeitos dos grupos experimental e controle.

As variáveis potenciais que não são de interesse direto para o pesquisador, mas que, mesmo assim, podem ser fontes de confusão no experimento, são chamadas de *variáveis externas*. Mas não deixe o termo enganá-lo! Um experimento confundido por uma variável externa não é menos confundido do que se a variável confundidora fosse de considerável interesse inerente. Por exemplo, Evans e Donnerstein (1974) observaram que os estudantes que se oferecem como voluntários para pesquisas no começo do período acadêmico têm mais orientação acadêmica e são mais prováveis de ter um lócus interno de controle (i.e., enfatizam sua própria responsabilidade, em vez de fatores externos, por seus atos) do que estudantes que se oferecem mais adiante no período. Seus resultados sugerem que não seria sensato testar todos os participantes da condição experimental no começo do período e os participantes da condição de controle no final do período, pois isso poderia confundir a variável independente com características dos participantes (p.ex., lócus de controle, foco acadêmico).

A randomização em bloco controla variáveis externas, balanceando-as entre os grupos. Tudo de que se precisa é que blocos inteiros sejam testados a cada nível da variável externa. Por exemplo, se houvesse quatro experimentadores, blocos inteiros do projeto de blocos randomizados seriam designados a cada experimentador. Como cada bloco contém todas as condições do experimento, essa estratégia garante que cada condição seja testada por cada experimentador. Normalmente, designaríamos o mesmo número de blocos a cada experimentador, mas isso não é essencial. O essencial é que blocos inteiros sejam testados a cada nível da variável externa, que, nesse caso, envolve os quatro experimentadores. O balanceamento pode se tornar um pouco complexo quando existem diversas variáveis externas, mas um planejamento prévio cuidadoso pode evitar a confusão com esses fatores.

Perda de sujeitos Enfatizamos que a lógica do desenho com grupos aleatórios exige que os grupos em um experimento difiram apenas por causa dos níveis da variável independente. Vimos que formar grupos comparáveis de sujeitos no começo de um experimento é outra característica essencial do desenho de grupos aleatórios. É igualmente importante que os grupos sejam comparáveis ao final do experimento, com exceção da variável independente. Quando os sujeitos começam um experimento mas não terminam, a validade interna do experimento pode ser ameaçada. É importante distinguir as duas maneiras em que os sujeitos podem não concluir um experimento: a perda mecânica de sujeitos e a perda seletiva de sujeitos.

A **perda mecânica de sujeitos** ocorre quando os sujeitos não concluem o experimento por falha de um equipamento (nesse caso, o experimentador é considerado parte do equipamento). A perda mecânica de sujeitos pode ocorrer se um computador estragar, ou se o experimentador ler as instruções incorretas, ou se alguém interromper uma seção experimental inadvertidamente. A perda mecânica é um problema menos crítico do que a perda seletiva, pois é improvável que a perda em si esteja relacionada com alguma característica do sujeito. Desse modo, a perda mecânica não deve levar a diferenças sistemáticas entre as características dos sujeitos que concluem o experimento nas suas diferentes condições. A perda mecânica também pode ser compreendida como o resultado de eventos fortuitos que devem ocorrer igualmente entre os grupos. Assim, a validade interna não costuma ser ameaçada quando é preciso excluir sujeitos do experimento por perda mecânica. Quando ocorre perda mecânica de sujeitos, ela deve ser documentada, devendo-se registrar o nome ou número do sujeito excluído e a razão para a perda. O sujeito perdido deve ser substituído pelo próximo sujeito testado.

A perda seletiva de sujeitos é uma questão muito mais séria. A **perda seletiva de sujeitos** ocorre (1) quando os sujeitos são perdidos de maneira diferencial entre as condições do experimento; (2) quando alguma característica do sujeito é responsável pela perda; e (3) quando essa característica do sujeito está relacionada com a variável dependente usada para avaliar o resultado do estudo. A perda seletiva de sujeitos destrói os grupos comparáveis que são essenciais para a lógica do desenho com grupos aleatórios e, assim, podem impossibilitar a interpretação do experimento.

Podemos ilustrar os problemas associados à perda seletiva de sujeitos considerando um exemplo fictício, mas realista. Suponhamos que os diretores de uma academia de ginástica decidam testar a efetividade de um programa de ginástica de um mês. Oitenta pessoas se apresentam como voluntários para o experimento, e são divididas aleatoriamente, 40 para cada grupo. A designação aleatória a condições cria grupos comparáveis no começo do experimento, balanceando características dos indivíduos, como peso, nível de preparo físico, motivação, etc., entre os dois grupos. Os membros

☑ Exercício II

Neste exercício, você precisará de um baralho. Deixe de lado o valete, o rei e a rainha e use as cartas de 1 a 10 (atribua o valor de 1 ao ás). Embaralhe bem as cartas.

Para ter uma noção de como a designação aleatória a condições funciona para criar grupos equivalentes, divida as cartas embaralhadas (randomizadas) em duas pilhas, cada uma com 20 cartas. Uma pilha representa os "sujeitos" designados aleatoriamente a uma condição experimental, e a segunda pilha representa sujeitos designados aleatoriamente a uma condição de controle. Suponhamos que o valor em cada carta indique o escore dos sujeitos (1-10) em uma medida de diferenças individuais, como a capacidade da memória.

1. Calcule um escore médio para os sujeitos em cada condição (pilha), somando o valor de cada carta e dividindo por 20. Os dois grupos são equivalentes em termos da sua capacidade média de memória?

Para entender os problemas associados à perda seletiva de sujeitos, suponha que os sujeitos com pouca capacidade de memória (valores de 1 e 2) são incapazes de completar o teste experimental e abandonam a condição experimental. Para simular isso, remova as cartas com valores de 1 e 2 da pilha que representa a sua condição experimental.

2. Calcule um novo escore médio para a pilha na condição experimental. Depois da perda seletiva de sujeitos, como se comparam os escores médios da capacidade de memória dos dois grupos? O que isso indica para a equivalência dos dois grupos formados inicialmente usando designação aleatória?
3. Para cada "sujeito" (carta) que abandonou o grupo experimental, remova uma carta comparável do grupo controle. Observe que você pode não ter combinações exatas, e pode ter que substituir um "1" por um "2" ou vice-versa. Calcule uma nova média para o grupo de controle. Esse procedimento restaura a equivalência inicial dos dois grupos?
4. Embaralhe as 40 cartas novamente e divida as cartas em quatro grupos. Calcule uma média para cada pilha de 10 cartas. Com menos "sujeitos" em cada grupo, a randomização (embaralhar) levou a grupos equivalentes?

do grupo controle apenas devem fazer um teste de preparo físico ao final do mês. Os do grupo experimental participam de um vigoroso programa de ginástica por um mês, antes de fazerem o teste. Suponhamos que todos os 38 participantes do controle compareçam ao teste ao final do mês, mas apenas 25 dos participantes experimentais continuem o rigoroso programa pelo mês inteiro. Suponhamos também que o escore médio de forma física para as 25 pessoas restantes no grupo experimental seja significativamente maior do que o escore médio das 40 pessoas do grupo controle. Os diretores da academia então fazem a seguinte afirmação: "um estudo científico mostrou que o nosso programa leva a uma melhor forma física".

Você acha que a afirmação da academia de ginástica é justificada? Não é. Esse estudo hipotético representa um exemplo clássico de perda seletiva de sujeitos, de modo que seus resultados não podem ser usados para corroborar a afirmação da academia. A perda ocorreu diferencialmente entre as condições, pois foram perdidos participantes principalmente do grupo experimental. O problema com a perda diferencial não é que os grupos tenham terminado com tamanhos diferentes. Os resultados teriam sido interpretáveis se 25 pessoas tivessem sido designadas aleatoriamente ao grupo experimental e 38 ao grupo controle e todos os indivíduos tivessem concluído o experimento. Ao contrário, a perda seletiva de sujeitos é um problema porque os 25 participantes experimentais que concluíram o programa de ginástica provavelmente não são comparáveis com os 38 participantes do controle. É provável que os 15 participantes experimentais que não conseguiram concluir o rigoroso programa estivessem [illegible] forma física (mesmo antes que o [illegible]rama começasse) do que os 25 participantes experimentais que concluíram o programa. A perda seletiva de sujeitos no grupo experimental arruinou os grupos comparáveis que foram formados por designação aleatória no começo do experimento. De fato, os escores finais de forma física dos 25 participantes experimentais poderiam ter sido maiores do que a média do grupo controle, mesmo que não tivessem participado do programa de ginástica, porque estavam em melhor forma quando começaram! Assim, a perda de sujeitos no experimento cumpre as outras duas condições para a perda seletiva de sujeitos. Ou seja, a perda provavelmente se deve a uma característica dos participantes – seu nível original de forma física – e essa característica é relevante para o resultado do estudo (ver Figura 6.3).

Se a perda seletiva de sujeitos não for identificada até a conclusão do experimento, pouco se pode fazer além de engolir a experiência de ter feito um experimento que não pode ser interpretado. Todavia, podem ser adotadas medidas quando os pesquisadores entendem antecipadamente que a perda seletiva pode ser um problema. Uma alternativa é administrar um pré-teste e triar sujeitos prováveis de ser perdidos. Por exemplo, no estudo sobre o programa de ginástica, podia ter sido aplicado um teste inicial da forma física, e apenas os participantes que tivessem um nível mínimo participariam do experimento. Triar os participantes desse modo envolveria um custo potencial. Os resultados do estudo provavelmente se aplicariam apenas a pessoas acima do nível mínimo de forma física. Talvez valesse a pena pagar esse custo, pois um estudo interpretável de generalização limitada ainda é preferível do que um estudo que não possa ser interpretado.

Existe outra abordagem preventiva que os pesquisadores podem usar ao en-

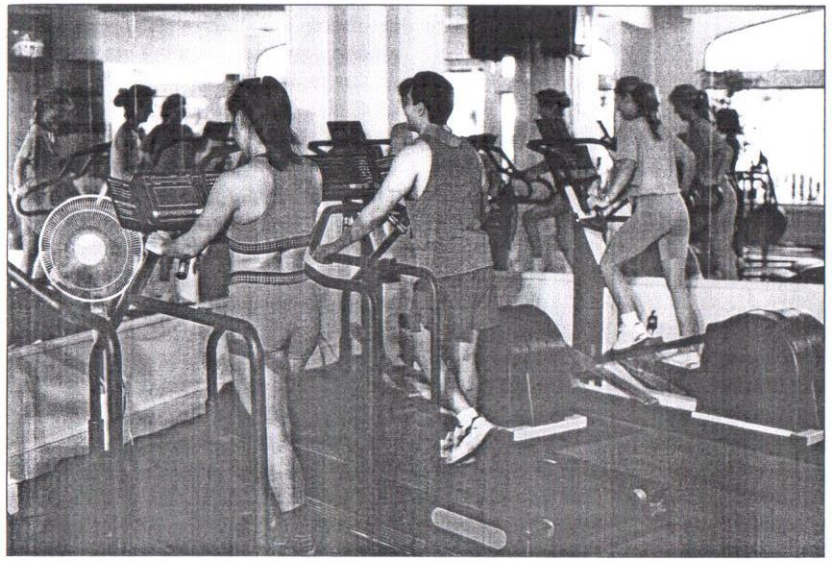

Figura 6.3 Muitas pessoas que começam um programa rigoroso de exercícios não o concluem. De certo modo, apenas os "mais aptos" sobrevivem, uma situação que pode causar problemas de interpretação em comparações entre tipos diferentes de programas de ginástica.

frentarem a possibilidade de perda seletiva de sujeitos. Os pesquisadores podem administrar um pré-teste a todos os sujeitos, mas depois simplesmente designar os participantes aleatoriamente às condições. Então, se um sujeito for perdido no grupo experimental, pode-se excluir um sujeito com um pré-teste comparável do grupo controle. De certo modo, essa abordagem visa restaurar a comparabilidade inicial dos grupos. Os pesquisadores devem ser capazes de prever possíveis fatores que possam levar à perda seletiva de sujeitos, e devem garantir que seu pré-teste avalie esses fatores.

Experimentos de controle com placebo e duplos-cegos O último desafio à validade interna que descreveremos ocorre por causa das expectativas dos participantes e experimentadores. As características de demanda representam uma fonte possível de viés devido às expectativas dos participantes (Orne, 1962). As *características de demanda* se referem às pistas e outras informações que os participantes usam para orientar seu comportamento em um estudo psicológico (ver Capítulo 4). Por exemplo, os participantes da pesquisa que sabem que tomarão álcool em um experimento podem esperar certos efeitos, como relaxamento ou tontura. Portanto, podem agir de maneira coerente com essas expectativas, em vez de responderem aos efeitos reais do álcool. Também podem surgir vieses potenciais por causa das expectativas dos experimentadores. O termo geral usado para descrever esses vieses é **efeitos do experimentador** (Rosenthal, 1963, 1994a). Os efeitos do experimentador podem ser uma fonte de confusão, se os experimentadores tratarem os sujeitos de maneiras distintas nos diferentes grupos do experimento, e distintas das exigidas para implementar a variável independente. Em um experimento envolvendo beber álcool, por exemplo, os efeitos do experimentador podem ocorrer se os experimentadores lerem as instruções de forma mais lenta para sujeitos que tiverem bebido do que para os que não beberem. Os efeitos do experimentador também podem ocorrer quando os experimentadores fazem observações tendenciosas, baseadas no tratamento que o sujeito recebeu. Por exemplo, poderia haver observações tendenciosas no estudo do álcool se os experimentadores fossem mais prováveis de observar movimentos motores inusitados ou fala arrastada entre os "bêbados" (pois "esperam" que quem bebe aja desse modo). (Ver a discussão sobre os efeitos da expectativa no Capítulo 4.)

Os pesquisadores nunca conseguem eliminar completamente os problemas das características de demanda e efeitos do experimentador, mas existem desenhos de pesquisa especiais que controlam esses problemas. Os pesquisadores usam um **grupo controle com placebo** como forma de controlar as características de demanda. Um *placebo* (da palavra latina que significa "devo agradar") é uma substância que parece como uma droga ou outra substância ativa, mas que na verdade é uma substância inerte, ou inativa. Algumas pesquisas até indicam que mesmo o placebo pode ter efeitos terapêuticos, com base em expectativas dos participantes para um efeito de uma "droga" (p.ex., Kirsch e Sapirstein, 1998). Os pesquisadores testam a eficácia de um tratamento proposto, comparando-o a um placebo. Ambos os grupos têm a mesma "consciência" de tomarem uma droga e, portanto, expectativas semelhantes para um efeito terapêutico. Ou seja, as características de demanda são semelhantes para os grupos – os participantes em ambos os grupos esperam sentir os efeitos de uma droga. Quaisquer diferenças entre os grupos experimentais e o grupo controle com placebo podem ser atribuídas legitimamente ao efeito real da droga tomada pelos sujeitos experimentais, e não a suas expectativas por tomarem a droga.

O uso de grupos controle com placebo em combinação com um procedimento duplo-cego pode controlar as características de demanda e os efeitos do experimentador. Em um **procedimento duplo-cego**, o participante e o observador estão cegos (desconhecem) ao tratamento que está sendo administrado. Em um experimento testando a eficácia de um tratamento farmacológico, seriam necessários dois pesquisadores para fazer o procedimento duplo-cego. O primeiro pesquisador prepararia as cápsulas com a droga e codificaria cada cápsula de algum modo; o segundo pesquisador distribuiria as drogas aos participantes, registrando o código para cada droga quando fosse dada a um indivíduo. Esse procedimento garante que haja um registro de qual droga cada pessoa recebeu, mas o participante e o experimentador que administra as drogas (e observa os efeitos) não sabem qual tratamento o sujeito recebeu. Assim, as expectativas do experimentador sobre os efeitos do tratamento são controladas, pois o pesquisador que faz as observações não está ciente de quem recebeu o tratamento e quem recebeu o placebo. De maneira semelhante, as características de demanda são controladas, pois os participantes permanecem sem saber se receberam a droga ou o placebo.

Os experimentos que envolvem grupos controle com placebo são uma ferramenta de pesquisa valiosa para avaliar a eficácia de um tratamento, enquanto controlam as características de demanda. Todavia, o uso de grupos controle com placebo suscita questões éticas especiais. Os benefícios do conhecimento adquirido com o uso de placebos devem ser avaliados à luz dos riscos envolvidos quando sujeitos de pesquisa que esperam tomar uma droga recebem um placebo em seu lugar. Geralmente, a ética desse procedimento é abordada no procedimento de consentimento informado, antes do começo do experimento. Os participantes são informados de que podem receber uma droga ou um placebo. Somente indivíduos que consentirem em tomar o placebo e a droga participam da pesquisa. Se a droga experimental se mostrar efetiva, os pesquisadores são eticamente obrigados a oferecer o tratamento para os participantes da condição do placebo.

Análise e interpretação de resultados experimentais

O papel da análise de dados em experimentos

- A análise de dados e a estatística desempenham um papel crítico na capacidade dos pesquisadores de afirmar que uma variável independente teve algum efeito sobre o comportamento.
- A melhor maneira de determinar se os resultados de um experimento são confiáveis é fazer uma replicação do experimento.

Um bom experimento, como ocorre com toda a boa pesquisa, começa com uma boa pergunta de pesquisa. Descrevemos como os pesquisadores usam as técnicas de controle para desenhar e implementar um experimento que lhes permita reunir evidências interpretáveis para responder sua pergunta de pesquisa. Todavia, apenas fazer um bom experimento não é suficiente. Os pesquisadores também devem apresentar as evidências de um modo convincente para demonstrar que seus dados corroboram suas conclusões baseadas naquelas evidências. A análise de dados e a estatística desempenham um papel crítico na análise e interpretação de resultados experimentais.

Robert Abelson, em seu livro *Statistics as Principled Argument* (1995), sugere que o principal objetivo da análise de dados é determinar se as observações sustentam uma afirmação sobre o comportamento. Ou seja, podemos "provar nosso argumento" para a conclusão baseada nas evidências reunidas em um experimento? Nos Capítulos 11 e 12, apresentamos uma descrição mais complexa de como os pesquisadores usam análise de dados e estatística. Aqui, iremos introduzir os conceitos centrais de análise de dados que se aplicam à interpretação dos resultados de experimentos. Antes, porém, queremos mencionar uma maneira muito importante pela qual os pesquisadores fazem seu argumento relacionado com os resultados de sua pesquisa.

A melhor maneira de determinar se os resultados obtidos em um experimento são fidedignos (consistentes) é replicar o experimento e ver se o mesmo resultado é obtido. A **replicação** significa repetir os procedimentos usados em um determinado experimento para determinar se os mesmos resultados serão obtidos uma segunda vez. Como você pode imaginar, uma replicação exata é quase impossível de executar. Os sujeitos testados na replicação serão diferentes dos testados no estudo original; as salas de teste e os experimentadores também podem ser diferentes. Entretanto, a replicação ainda é a melhor maneira de determinar se o resultado de uma pesquisa é fidedigno. Contudo, se exigíssemos que a fidedignidade de cada experimento fosse estabelecida por replicação, o processo seria incômodo e ineficiente. Os participantes de experimentos são um recurso escasso, e fazer uma replicação significa que estaremos deixando de fazer um experimento para fazer perguntas novas e diferentes sobre o comportamento. A análise de dados e a estatística proporcionam uma alternativa à replicação para os pesquisadores determinarem se os resultados de um único experimento são fidedignos e podem ser usados para fazer uma afirmação sobre o efeito que uma variável independente sobre o comportamento.

Dica de estatística

A análise dos dados de um experimento envolve três estágios: (1) conhecer os dados, (2) sintetizar os dados e (3) confirmar o que os dados revelam. No primeiro estágio, tentamos descobrir o que está acontecendo no conjunto de dados, procuramos erros e nos certificamos de que os dados fazem sentido. No segundo estágio, usamos estatísticas descritivas e demonstrações gráficas para sintetizar o que se descobriu. No terceiro estágio, buscamos evidências para o que os dados nos dizem sobre o comportamento. Neste estágio, tiramos nossas conclusões sobre os dados usando técnicas estatísticas variadas.

Nas próximas seções, apresentamos aqui nas uma introdução sucinta a esses estágios da análise de dados. Uma introdução mais completa pode ser encontrada nos Capítulos 11 e 12 (ver especialmente o Quadro 11.1). Esses capítulos serão particularmente importantes se você precisar ler e interpretar os resultados de um experimento de psicologia publicado em uma revista científica ou se fizer o seu próprio experimento de psicologia.

Ilustraremos o processo de análise de dados analisando os resultados de um experimento que investigou os efeitos de recompensas e punições enquanto os participantes jogavam *videogames* violentos. Carnagey e Anderson (2005) observaram que uma grande *corpus* de pesquisas demonstra que jogar *videogames* violentos aumenta as emoções, cognições e comportamentos agressivos. Eles questionaram, contudo, se os efeitos de *videogames* violentos seriam diferentes quando os jogadores fossem *punidos* por ações violentas nos jogos em comparação com quando as mesmas ações são *recompensadas* (como na maioria dos *videogames*). Uma hipótese postulada por Carnagey e Anderson foi que quando as ações violentas nos *videogames* fossem punidas, os jogadores seriam menos agressivos. Outra hipótese, contudo, dizia que, quando punidos por seus atos violentos, os jogadores ficariam frustrados e, portanto, mais agressivos.

Nos estudos de Carnagey e Anderson, estudantes de graduação jogaram três versões do mesmo *videogames* com uma corrida de carros competitiva ("Carmageddon 2") em um ambiente laboratorial. Na condição de recompensa, os participantes foram recompensados (ganharam pontos) por matarem pedestres e o oponente na corrida (essa é a versão do jogo inalterada). Na condição de punição, o *videogames* foi alterado de maneira que os participantes perdiam pontos por matar ou bater nos oponentes. Em uma terceira condição, o jogo foi alterado para ser não violento, e os participantes ganhavam pontos por passarem por pontos de pedestre a medida que corriam ao redor da pista (todos os pedestres foram retirados e os oponentes foram programados para serem passivos).

Carnagey e Anderson (2005) publicaram os resultados de três experimentos, nos quais os sujeitos foram designados aleatoriamente para jogar uma das três versões do *videogames*. As principais variáveis dependentes eram medidas das emoções hostis dos participantes (Experimento 1), pensamento agressivo (Experimento 2) e comportamentos agressivos (Experimento 3). Entre os três estudos, os participantes que foram recompensados por atos violentos no *videogames* tiveram níveis maiores de emoções, cognições e comportamentos agressivos, comparados com as condições de jogo com punição e não violenta. Punir atos agressivos no *videogames* fez os participantes sentirem mais emoções hostis (semelhante à condição de recompensa) em relação ao jogo não violento, mas não os fez ter mais cognições e comportamentos agressivos.

Para ilustrar o processo de análise de dados, analisaremos de forma mais minuciosa os resultados de Carnagey e Anderson para cognições agressivas (Experimento 2). Depois de jogarem um dos três *videogames*, os participantes fizeram um teste com fragmentos de palavras, no qual deviam completar o maior número de palavras (de 98) que conseguissem em cinco minutos. Metade dos fragmentos de palavras tinha possibilidades agressivas. Por exemplo, o fragmento "K I _ _" podia ser completado como *kiss* (beijo) ou *kill* (matar) (ou outras possibilidades). A cognição agressiva foi definida operacionalmente como a proporção de fragmentos de palavras que um participante completasse com palavras agressivas. Por exemplo, se um participante completasse 60 dos fragmentos de palavras em cinco minutos e 12 delas expressassem conteúdo agressivo, seu escore de cognição agressiva seria 0,20 (i.e., 12/60 = 0,20).

Descrevendo os resultados

- As duas estatísticas descritivas mais comuns que são usadas para sintetizar os resultados de experimentos são a média e o desvio padrão.
- As medidas do tamanho do efeito indicam a intensidade da relação entre as variáveis independentes e dependentes, e não são afetadas pelo tamanho da amostra.
- Uma medida comum do tamanho do efeito, *d*, analisa a diferença entre duas médias grupais, em relação à variabilidade média no experimento.
- A meta-análise usa medidas do tamanho do efeito para sintetizar os resultados de muitos experimentos que investigam a mesma variável independente ou dependente.

A análise de dados deve começar com uma inspeção minuciosa do conjunto de dados, com especial atenção a erros possíveis ou dados anômalos. Técnicas para inspecionar os dados ("conhecer os dados") são descritas no Capítulo 11. O próximo passo é descrever o que se encontrou. Nesse estágio, o pesquisador quer saber "o que aconteceu no experimento?". Para começar a

Tabela 6.1 Médias das cognições agressivas, desvios padrão e intervalos de confiança para as três condições do experimento com o *videogame*

Versão do *videogame*	Média	DP	Intervalo de confiança de 0,95*
Recompensa	0,210	0,066	0,186-0,234
Punição	0,175	0,046	0,151-0,199
Não violento	0,157	0,050	0,133-0,181

*Intervalos de confiança estimados com base em dados publicados em Carnagey e Anderson (2005).

responder essa pergunta, os pesquisadores usam *estatística descritiva*. As duas estatísticas descritivas mais comuns são a média (uma medida da tendência central) e o desvio padrão (uma medida da variabilidade). As médias e desvios padrão para a cognição agressiva no experimento do *videogame* são apresentados na Tabela 6.1. A média mostra que a cognição agressiva foi maior na condição de recompensa (0,210) e menor na condição não violenta (0,157). A cognição agressiva na condição de punição (0,175) ficou entre as condições não violenta e de recompensa. Podemos observar que, para participantes da condição de recompensa, aproximadamente uma em cada cinco palavras foi completada com conteúdo agressivo (lembre, porém, que apenas metade dos fragmentos de palavras tinha possibilidades agressivas).

Em um experimento conduzido adequadamente, o desvio padrão de cada grupo deve refletir apenas as diferenças individuais entre os sujeitos que foram designados aleatoriamente àquele grupo. Os sujeitos em cada grupo devem ser tratados do mesmo modo, e o nível da variável independente a que foram designados deve ser implementado da mesma forma para cada sujeito no grupo. Os desvios padrão mostrados na Tabela 6.1 indicam que houve variação ao redor da média em cada grupo e que a variação foi aproximadamente a mesma em todos os três grupos.

Uma pergunta importante que os pesquisadores fazem ao descrever os resultados de um experimento é sobre o tamanho do efeito que a variável independente teve sobre a variável dependente. As medidas do **tamanho do efeito** podem ser usadas para responder essa pergunta, pois indicam a intensidade da relação entre as variáveis independentes e dependentes. Uma vantagem das medidas do tamanho do efeito é que elas não são influenciadas pelo tamanho das amostras testadas no experimento. As medidas do tamanho do efeito levam em conta mais do que a diferença média entre duas condições de um experimento. A diferença média entre dois grupos sempre é *relativa* à variabilidade média nos escores dos participantes. Uma medida do tamanho do efeito usada com frequência é o ***d* de Cohen**. Cohen (1992) desenvolveu procedimentos que hoje são aceitos amplamente. Ele sugeriu que valores de *d* de 0,20, 0,50 e 0,80 representam efeitos pequenos, médios e grandes da variável independente, respectivamente.

Podemos ilustrar o uso do *d* de Cohen como medida do tamanho do efeito comparando duas condições no experimento do *videogame*, a condição de recompensa e a condição não violenta. O valor de *d* é 0,83, com base na diferença entre a cognição agressiva média na condição de recompensa (0,210) e a condição não violenta (0,157). Esse valor de *d* nos permite dizer que a variável independente do *videogame*, recompensa *versus* não violento, teve um efeito grande sobre a cognição agressiva nessas duas condições. As medidas do tamanho do efeito fornecem informações valiosas para os pesquisadores descreverem os resultados de um experimento.

Dica de estatística

As medidas da tendência central e da variabilidade, assim como do tamanho do efeito, são descritas nos Capítulos 11 e 12. Nesses capítulos, apresentamos os passos matemáticos para essas medidas e discutimos sua interpretação. Muitas medidas diferentes do tamanho do efeito são encontradas na literatura em psicologia. Além do *d* de Cohen, por exemplo, uma medida popular da magnitude do efeito é o eta quadrado, que é uma medida da intensidade da associação entre as variáveis independentes e dependentes (ver o Capítulo 12). Ou seja, o eta quadrado estima a proporção da variância total explicada pelo efeito da variável independente sobre a variável dependente. As medidas do tamanho do efeito são mais úteis ao se compararem os valores numéricos de uma medida de dois ou mais estudos ou quando se calculam médias de medidas de estudos, como em uma meta-análise (ver a seguir).

Os pesquisadores também usam medidas do tamanho do efeito em um procedimento chamado de **meta-análise**. A meta-análise é uma técnica estatística usada para sintetizar os tamanhos dos efeitos de vários experimentos independentes que investigam a mesma variável independente ou dependente. De um modo geral, a qualidade metodológica dos experimentos incluídos na meta-análise determinará o seu valor final (ver Judd, Smith e Kidder, 1991). As meta-análises são usadas para responder perguntas como: existem diferenças de gênero na conformidade? Quais são os efeitos do tamanho da classe no desempenho acadêmico? A terapia cognitiva é efetiva no tra-

Quadro 6.2

EXEMPLO DE META-ANÁLISE: "PSICOTERAPIAS BASEADAS EM EVIDÊNCIAS PARA JOVENS *VERSUS* TRATAMENTO CLÍNICO USUAL"

Weisz, Jensen-Doss e Hawley (2006) usaram uma meta-análise para sintetizar os resultados de 32 estudos sobre psicoterapias com jovens, comparando os efeitos de "tratamentos baseados em evidências" e "tratamento usual". Um tratamento baseado em evidências é aquele que tem amparo empírico – ou seja, que, na prática clínica, demonstrou ajudar indivíduos. Embora pareça óbvio que os tratamentos baseados em evidências devam ser amplamente utilizados na prática clínica por causa desse amparo empírico, muitos terapeutas argumentam que esses tratamentos não seriam efetivos em contextos clínicos usuais. Os tratamentos baseados em evidências são estruturados e exigem que os terapeutas sigam um manual de tratamento. Alguns clínicos argumentam que esses tratamentos são rígidos e inflexíveis, não podendo ser individualizados conforme as necessidades dos clientes. Além disso, os oponentes dos tratamentos baseados em evidências dizem que os estudos empíricos que indicam a sua efetividade geralmente envolvem clientes com problemas menos graves ou complicados do que os observados na prática clínica usual. Esses argumentos sugerem que o tratamento usual, na forma de psicoterapia, aconselhamento ou manejo de caso, conduzido regularmente por profissionais da saúde mental, seria mais capaz de satisfazer as necessidades dos clientes atendidos normalmente em ambientes na comunidade.

Weisz e seus colegas usaram meta-análise para comparar diretamente os resultados associados aos tratamentos baseados em evidências e o tratamento usual. Entre 32 estudos comparando os dois modelos, o tamanho do efeito médio foi de 0,30. Assim, os jovens tratados com um tratamento baseado em evidências foram mais beneficiados, em média, do que os tratados da maneira usual. O valor de 0,30 está entre os critérios de Cohen (1988) para efeitos pequenos e médios. Esse tamanho de efeito representa a diferença entre os dois tipos de tratamentos, e não o efeito da psicoterapia em si. Weisz e colaboradores observam que, quando tratamentos baseados em evidências são comparados com grupos de controle sem tratamento (p.ex., lista de espera), seus tamanhos de efeito geralmente variam de 0,50 a 0,80 (efeitos médios a grandes). Em outras análises, os autores agruparam estudos segundo fatores como a gravidade e a complexidade dos problemas tratados, ambientes de tratamento e características dos terapeutas. Essas análises visavam determinar se as preocupações apontadas pelos críticos de tratamentos baseados em evidências justificavam o uso continuado do tratamento usual. Weisz e seus colegas observaram que agrupar estudos segundo esses diversos fatores não influenciou o resultado geral, de que os tratamentos baseados em evidências são melhores do que o tratamento usual.

Essa meta-análise permite que os psicólogos defendam, com mais confiança, um princípio psicológico geral relacionado com a psicoterapia: os tratamentos baseados em evidências proporcionam resultados melhores para os jovens do que o tratamento usual.

tamento da depressão? O Quadro 6.2 descreve uma meta-análise de estudos sobre a psicoterapia efetiva para jovens com transtornos psicológicos. Os resultados de experimentos individuais, não importa o quão bem feitos, muitas vezes não são suficientes para fornecer respostas para perguntas sobre questões gerais importantes. Devemos considerar um *corpus* de literatura (i.e., muitos experimentos) relacionado a cada questão. (Ver Hunt, 1997, para uma introdução boa e fácil de ler à meta-análise.)

As meta-análises nos permitem tirar conclusões mais firmes sobre os princípios da psicologia, pois essas conclusões somente emergem após se analisarem os resultados de muitos experimentos individuais. Essas análises são um modo eficiente e efetivo de sintetizar os resultados de grandes números de experimentos usando medidas do tamanho do efeito.

Confirmando o que os resultados revelam

- Os pesquisadores usam estatísticas inferenciais para determinar se uma variável independente tem um efeito fidedigno sobre uma variável dependente.
- Dois métodos de fazer inferências baseadas em dados amostrais são testar uma hipótese nula e intervalos de confiança.
- Os pesquisadores usam o teste da hipótese nula para determinar se diferenças médias entre grupos em um experimento são maiores do que as diferenças esperadas simplesmente pela variação do erro experimental.
- Um resultado estatisticamente significativo é aquele que tem uma probabilidade pequena de ocorrer se a hipótese nula for verdadeira.
- Os pesquisadores determinam se uma variável independente teve um efeito sobre o comportamento analisando se existe sobreposição entre os intervalos de confiança para as diferentes amostras do experimento. O grau de sobreposição informa se a média amostral estima a mesma média populacional ou médias populacionais diferentes.

Talvez a afirmação mais básica que os pesquisadores desejam fazer quando realizam um experimento seja que a variável independente teve um efeito sobre a variável dependente. Outra maneira de formular essa afirmação é dizer que os pesquisadores querem confirmar que a variável independente *produziu uma diferença no comportamento*. As estatísticas descritivas, por si só, não são evidências suficientes para confirmar essa afirmação básica.

Para confirmar se a variável independente teve efeito em um experimento, os pesquisadores usam *estatística inferencial*. Eles precisam usar estatística inferencial por causa da natureza do controle proporcionado pela designação aleatória em experimentos. Como descrevemos anteriormente, a designação aleatória não *elimina* as diferenças individuais entre os sujeitos. A designação aleatória simplesmente *equilibra* ou faz a média das diferenças individuais entre os sujeitos dos grupos do experimento. A variação assistemática (i.e., aleatória) decorrente das diferenças entre os sujeitos de cada grupo se chama *variação do erro*. A presença dessa variação representa um problema potencial, pois a média dos diferentes grupos do experimento pode diferir simplesmente por causa da variação do erro, e não porque a variável independente teve efeito. Assim, por si só, os resultados médios do mais bem controlado experimento não permitem concluir definitivamente se a variável independente produziu uma diferença no comportamento. A estatística inferencial permite que os pesquisadores testem se as diferenças na média grupal se devem a um efeito da variável independente, e não apenas ao acaso (variação do erro). Os pesquisadores usam dois tipos de estatística inferencial para decidir se uma variável independente teve um efeito: teste da hipótese nula e intervalos de confiança.

Dica de estatística

Sabemos que pode ser frustrante descobrir que os resultados do mais bem controlado experimento muitas vezes não permitem concluir definitivamente se a variável independente gerou uma diferença no comportamento. Em outras palavras, o que você aprendeu até aqui sobre métodos de pesquisa não é suficiente! Infelizmente, mesmo com as ferramentas da análise de dados, não podemos lhe dar um modo de tirar conclusões *definitivas* sobre o que produziu uma diferença no comportamento. Porém, o que podemos lhe dar é um modo (na verdade, vários modos) de fazer a melhor afirmação possível sobre o que produziu a diferença. A conclusão se baseará em uma *probabilidade* – ou seja, uma probabilidade que lhe ajudará a decidir se o seu efeito se deve ou não apenas ao acaso. É fácil se perder nas complexidades do teste da hipótese nula e dos intervalos de confiança, mas tenha em mente os dois pontos críticos seguintes:

Antes de mais nada, as diferenças no comportamento podem surgir simplesmente por acaso (chamadas de *variação do erro*). O que você quer saber é: qual é a probabilidade de que a diferença que observou se deva apenas ao acaso (e não ao efeito da sua variável independente?). Na verdade, o que você realmente gostaria de saber é a probabilidade de que a sua variável independente cause um efeito. Todavia, não podemos responder essas perguntas usando inferência estatística. Como você verá, a inferência estatística é indireta (ver, por exemplo, o Quadro 12.1 no Capítulo 12).

Em segundo lugar, os dados que você coletou representam *amostras* de uma população; porém, de certo modo, são as *populações*, e não as amostras, que realmente importam. (Se o importante fosse as médias amostrais, você poderia apenas olhar as médias amostrais para ver se eram diferentes.) O desempenho médio das amostras nas várias condições do seu experimento fornece estimativas que são usadas para *inferir* a média da população. Quando faz afirmações de inferência estatística, você está usando a média amostral para tirar conclusões (fazer inferências) sobre as diferenças entre médias populacionais. Mais uma vez, sugerimos o Capítulo 12 para uma discussão mais complexa sobre essas questões.

Teste de significância da hipótese nula Os pesquisadores costumam utilizar um **teste de significância da hipótese nula** para verificar se uma variável independente teve um efeito em um experimento. O teste de significância da hipótese nula começa com a premissa de que a variável independente *não* teve efeito. Se pressupomos que a hipótese nula é verdadeira, podemos usar a teoria da probabilidade para determinar a probabilidade de a diferença que observamos em nosso experimento ocorrer apenas "por acaso". *Um resultado* **estatisticamente significativo** *é aquele que tem uma probabilidade apenas pequena de ocorrer se a hipótese nula for verdadeira.* Um resultado estatisticamente significativo significa apenas que a diferença que obtivemos em nosso experimento é maior do que seria de esperar se apenas a variação do erro (i.e., o acaso) fosse responsável pelo resultado.

Geralmente, expressa-se o resultado de um experimento em termos das diferenças entre as médias para as condições do experimento. Como sabemos a probabilidade do resultado obtido no experimento? Os pesquisadores costumam usar testes de estatística inferencial, como o teste t ou o teste F. O teste t é usado quando a variável independente tem dois níveis, e o teste F é usado quando ela tem três ou mais níveis. Cada valor de um teste t ou F tem uma probabilidade associada a ele quando se considera a hipótese nula, que pode ser determinada calculando-se o valor da estatística do teste.

Pressupondo que a hipótese nula seja verdadeira, quão pequena deve ser a probabilidade do nosso resultado para que seja estatisticamente significativa? Os cientistas tendem a concordar que resultados com probabilidades (p) de menos de 5 vezes em 100 (ou $p < 0{,}05$) são considerados estatisticamente significativos. A probabilidade que os pesquisadores usam para decidir se um resultado é estatisticamente significativo se chama *nível de significância*. O nível de significância é indicado pela letra grega alfa (α).

Agora, podemos ilustrar o procedimento do teste da hipótese nula para analisar o

experimento do *videogame* que descrevemos anteriormente (ver Tabela 6.1, p. 213). A primeira pergunta de pesquisa que faríamos é se a variável independente da versão do *videogame* tem algum efeito *geral*. Ou seja, a cognição agressiva difere em função das três versões do *videogame*? A hipótese nula para esse teste geral é de que não existe diferença entre a média populacional representada pelas médias das condições experimentais (lembre-se de que a hipótese nula pressupõe que a variável independente não tem efeito). O valor de p para o teste F calculado para o efeito da versão do *videogame* foi menor do que o nível de significância de 0,05; assim, o efeito geral da variável do *videogame* foi estatisticamente significativo. Para interpretar esse resultado, deveríamos nos reportar à estatística descritiva para esse experimento na Tabela 6.1, onde vemos que a cognição agressiva média para as três condições do *videogame* era diferente. Por exemplo, a cognição agressiva foi maior com o *videogame* com recompensa (0,210) e menor com o *videogame* não violento (0,157). O resultado estatisticamente significativo do teste F nos permite afirmar que a versão do *videogame* gerou uma diferença singular na cognição agressiva.

Os pesquisadores buscam fazer afirmações mais específicas a respeito dos efeitos de variáveis independentes sobre o comportamento do que apenas dizer que a variável independente tem efeito. Os testes F das diferenças gerais entre as médias nos dizem que algo aconteceu no experimento, mas não falam muito sobre o que aconteceu de fato. Uma maneira de obter informações mais específicas sobre os efeitos das variáveis independentes, é usar intervalos de confiança.

Usando intervalos de confiança para analisar diferenças médias Os intervalos de confiança para cada um dos três grupos no experimento do *videogame* são apresentados na Tabela 6.1, na página 213. Um intervalo de confiança é associado a uma probabilidade (geralmente de 0,95) de que o intervalo contenha a média populacional verdadeira. A amplitude do intervalo nos diz o quanto a nossa estimativa é precisa (quanto menor, melhor). Os **intervalos de confiança** também podem ser usados para comparar diferenças entre duas médias populacionais. Podemos usar os intervalos de confiança de 0,95 apresentados na Tabela 6.1 para fazer perguntas específicas sobre os efeitos da versão do *videogame* sobre a cognição agressiva. Faz-se isso analisando se existe sobreposição entre os intervalos de confiança para os diferentes grupos. *Quando os intervalos de confiança não se sobrepõem podemos ter confiança de que as médias populacionais para os dois grupos diferem*. Por exemplo, observe que o intervalo de confiança para o grupo da recompensa é de 0,186 a 0,234. Isso indica que existe uma probabilidade de 0,95 de que o intervalo de 0,186 a 0,234 contenha a média populacional para cognição agressiva na condição de recompensa (lembre que a média amostral de 0,210 apenas *estima* a média populacional). O intervalo de confiança para o grupo não violento é de 0,133 a 0,181. Esse intervalo de confiança não se sobrepõe com o intervalo de confiança do grupo da recompensa (i.e., o limite superior de 0,181 para o grupo não violento é menor que o limite inferior de 0,186 para o grupo da recompensa). Com essa evidência, podemos afirmar que a cognição agressiva na condição da recompensa foi maior do que a cognição agressiva na condição do *videogame* não violento.

Todavia, quando comparamos os intervalos de confiança do grupo do *videogame* com recompensa (0,186-0,234) e do grupo com punição (0,151-0,199), chegamos a uma conclusão diferente. Os intervalos de confiança para esses grupos se sobrepõem. Embora as médias amostrais de 0,210 e 0,175 difiram, não podemos concluir que a média populacional difira por causa da sobreposição dos intervalos de confiança. Podemos propor a seguinte regra básica para interpretar esse resultado: *se os intervalos se sobrepõem levemente, devemos reconhecer a nossa incerteza quanto à verdadeira diferença mé-*

dia, postergar qualquer juízo; se os intervalos se sobrepõem de modo que a média de um grupo esteja dentro do intervalo de outro grupo, podemos concluir que as médias populacionais não diferem. No experimento do *videogame*, a sobreposição é pequena e a média amostral de cada condição não fica dentro dos intervalos do outro grupo. Queremos saber se as populações diferem, mas tudo que podemos dizer é que não temos evidências suficientes para decidir por um ou outro lado. Nessa situação, devemos postergar qualquer decisão até o próximo experimento.

Dica de estatística

A lógica e os procedimentos computacionais para os intervalos de confiança e o teste t são encontrados no Capítulo 11. O teste F (em suas várias formas) é discutido no Capítulo 12.

O que a análise de dados não pode nos dizer

Já fizemos alusão a algo que nossa análise de dados não pode nos dizer. Mesmo que nosso experimento seja internamente válido e os resultados sejam estatisticamente significativos, não podemos dizer *com certeza* que a nossa variável independente teve efeito (ou que não teve). Devemos aprender a viver com afirmações probabilísticas. Os resultados da nossa análise de dados também não podem nos dizer se os resultados do nosso estudo têm valor prático ou mesmo se são significativos. É fácil fazer experimentos com perguntas de pesquisa triviais (ver Sternberg 1997, e Capítulo 1). Também é fácil (talvez fácil demais!) fazer um mau experimento. Os maus experimentos – ou seja, aqueles que carecem de validade interna – podem facilmente produzir resultados estatisticamente significativos e intervalos de confiança que não se sobreponham; todavia, o resultado não será interpretável.

Quando um resultado é estatisticamente significativo, concluímos que a variável independente teve um efeito sobre o comportamento. Ainda assim, como já vimos, nossa análise não nos possibilita ter certeza quanto à conclusão, mesmo que tenhamos chegado a essa conclusão "além de qualquer dúvida". Além disso, quando um resultado *não* é estatisticamente significativo, não podemos concluir com certeza que a variável independente *não* teve efeito. Tudo que podemos concluir é que não existem evidências suficientes para dizer que a variável independente produz um efeito. Determinar que uma variável independente não teve efeito pode ser ainda mais crucial na pesquisa aplicada. Por exemplo, será que um remédio genérico é tão efetivo quanto seu correlato de marca conhecida? Para responder essa pergunta de pesquisa, os pesquisadores muitas vezes tentam confirmar que não existem diferenças entre as drogas. Os padrões para experimentos que visam responder perguntas relacionadas com a ausência de diferenças entre condições são mais elevados do que para experimentos visando confirmar que uma variável independente tem efeito. Descreveremos esses padrões no Capítulo 12.

Como os pesquisadores baseiam-se em probabilidades para tirar conclusões sobre os efeitos de variáveis independentes, sempre existe a chance de se cometer um erro. Existem dois tipos de erros que podem ocorrer quando os pesquisadores usam estatística inferencial. Quando dizemos que um resultado é estatisticamente significativo e que a hipótese nula (não existe diferença) é realmente verdadeira, estamos cometendo um erro do Tipo I. Um *erro do Tipo I* é como um alarme falso – dizer que há um incêndio quando não há. Quando concluímos que temos evidências insuficientes para rejeitar a hipótese nula e ela, de fato, é falsa, estamos cometendo um *erro do Tipo II* (os erros do Tipo I e do Tipo II são descritos no Capítulo 12). Jamais cometeríamos algum desses erros se pudéssemos saber ao certo se a hipótese nula é verdadeira ou falsa. Mesmo sabendo da possibilidade de que a análise de dados pode levar a decisões incorretas, também devemos lembrar que a análise

de dados pode e de fato muitas vezes leva a decisões corretas. O mais importante a lembrar para os pesquisadores é que a estatística inferencial jamais pode substituir a replicação como teste final da fidedignidade do resultado de um experimento.

Estabelecendo a validade externa de resultados experimentais

- Os resultados de um experimento têm validade externa quando podem ser aplicados a outros indivíduos, situações e condições além do escopo do experimento específico.
- Em certas investigações (p.ex., teste de teorias), os pesquisadores podem decidir enfatizar a validade interna sobre a externa; outros pesquisadores podem preferir aumentar a validade externa usando amostragem ou replicação.
- Fazer experimentos de campo é um modo de os pesquisadores aumentarem a validade externa de suas pesquisas em situações no mundo real.
- A replicação parcial é um método usado para estabelecer a validade externa de resultados de pesquisa.
- Os pesquisadores muitas vezes buscam generalizar os resultados associados a relações conceituais entre variáveis, em vez de condições, manipulações, situações e amostras específicas.

Como você aprendeu no Capítulo 4, a *validade externa* se refere ao nível em que os resultados de um estudo podem ser generalizados para indivíduos, situações e condições além do escopo do estudo específico. Uma crítica frequente a experimentos muito controlados é que eles não possuem validade externa; ou seja, os resultados observados em um experimento laboratorial controlado podem descrever o que ocorre apenas naquela situação específica, com as condições específicas que foram testadas e com os indivíduos específicos que participaram. Considere novamente o experimento do *videogame*, no qual estudantes universitários jogaram um *videogame* com uma corrida de carros em um ambiente laboratorial. O ambiente laboratorial é idealmente adequado para usar procedimentos de controle que garantam a validade interna de um experimento. Mas será que esses resultados nos ajudam a entender a violência e a agressividade em uma situação natural? Quando existe um tipo diferente de exposição à violência? Quando as pessoas expostas à violência são idosas? Essas são questões ligadas à validade externa e suscitam uma questão mais geral. Se os resultados de experimentos laboratoriais são tão específicos, que benefícios eles trazem para a sociedade?

Uma resposta a essa pergunta é um pouco perturbadora, pelo menos inicialmente. Mook (1983) argumenta que, quando o propósito de um experimento é testar uma determinada hipótese derivada de uma teoria psicológica, a questão da validade externa dos resultados é irrelevante. É comum se fazer um experimento para determinar se os sujeitos *podem* ser induzidos a agir de um determinado modo. A questão de se os sujeitos *agem* desse modo em seu ambiente natural é secundária à pergunta levantada no experimento. A questão da validade externa dos experimentos não é nova, conforme reflete a seguinte afirmação de Riley (1962): "de um modo geral, os experimentos laboratoriais não são montados para imitar o caso mais típico encontrado na natureza. Ao contrário, elas visam responder uma pergunta específica de interesse para o experimentador" (p. 413).

É claro que os pesquisadores muitas vezes querem obter resultados que possam generalizar além dos limites do experimento em si. Para alcançar esse objetivo, os pesquisadores podem incluir, em seus experimentos, as características das situações para as quais gostariam de generalizá-los. Por exemplo, Ceci (1993) descreveu um programa de pesquisa que conduziu com seus colegas, sobre testemunhos oculares de crianças. O autor contou que seu programa

de pesquisa foi motivado em parte porque estudos prévios sobre esse tema não compreenderam todas as dimensões de uma situação *real* de testemunho. Ceci descreveu como seu programa de pesquisa incluiu fatores como entrevistas sugestivas múltiplas, períodos de retenção muito longos e recordações de experiências estressantes. A inclusão desses fatores tornou os experimentos mais representativos de situações que ocorrem quando crianças testemunham (ver Figura 6.4).

Todavia, Ceci (1993) também observou que permanecem diferenças importantes entre os experimentos e situações da vida real:

> Níveis elevados de estresse, agressões contra o corpo da vítima e a perda de controle são característicos de situações que motivam investigações forenses. Embora esses fatores estejam em jogo em alguns dos nossos outros estudos, jamais repetiremos de maneira experimental a natureza agressiva dos atos perpetrados em vítimas infantis, pois mesmo os estudos que mais se aproximam, como estudos médicos, são sancionados pelos pais e pela sociedade, ao contrário de agressões sexuais contra crianças. (p. 41-42)

Conforme revelam os comentários de Ceci, em certas situações, como aquelas que envolvem testemunhos oculares sobre atos vis, pode haver importantes limitações éticas ao estabelecimento da validade externa dos experimentos.

A validade externa de pesquisas costuma ser questionada por causa da natureza dos "sujeitos". Como se sabe, muitos estudos em psicologia envolvem estudan-

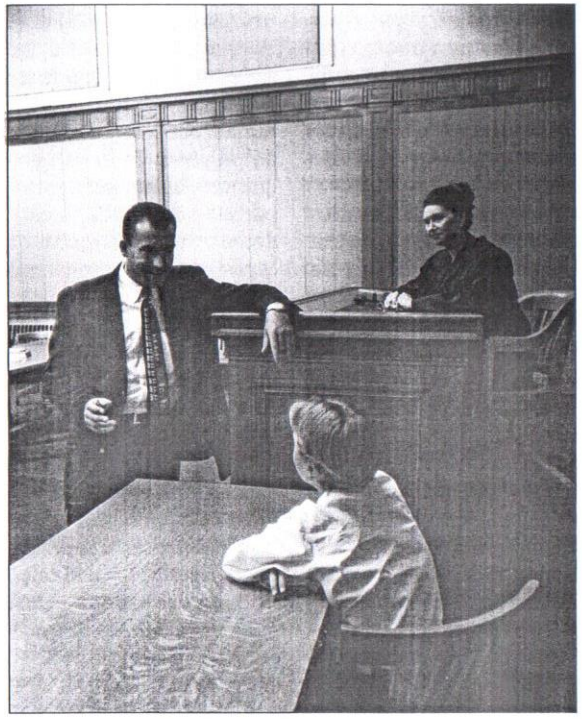

Figura 6.4 Em que nível os experimentos podem reproduzir as situações da vida real em que crianças testemunham em tribunais?

tes universitários que participam de experimentos como parte de sua disciplina de introdução à psicologia. Dawes (1991), entre outros, argumenta que os estudantes universitários são um grupo seleto, que nem sempre pode ser uma boa base para construir conclusões gerais sobre o comportamento e processos mentais humanos. De maneira semelhante, Sue (1999) afirma que a maior ênfase dos pesquisadores na validade interna sobre a externa diminui a atenção à representatividade das pessoas estudadas. Todavia, os psicólogos geralmente acreditam que seus resultados podem ser generalizados para populações além das especificamente testadas em suas pesquisas, e existe pouca razão para validar os resultados testando populações de minorias étnicas ou outras populações sub-representadas. Questões sobre a validade externa de resultados de pesquisa baseadas nas populações estudadas são especialmente importantes na pesquisa aplicada. Na pesquisa médica, por exemplo, os tratamentos efetivos para homens podem não ser efetivos para mulheres, e os tratamentos efetivos para adultos podem não ser efetivos para crianças.

Os *experimentos de campo*, que mencionamos rapidamente no Capítulo 4, são um modo de aumentar a validade externa de um estudo, podendo também gerar conhecimento prático. Por exemplo, para investigar as percepções das pessoas sobre os riscos, os participantes de dois experimentos de campo responderam a perguntas sobre riscos durante a pandemia da influenza H1N1 em 2009 (Lee, Schwartz, Taubman e Hou, 2010). O primeiro experimento foi realizado em um *campus* universitário, e o segundo foi realizado em *shopping centers* e perto da área comercial do centro da cidade. Os indivíduos que concordaram em participar foram designados aleatoriamente a uma condição experimental, na qual o cúmplice espirrava e tossia antes da administração de um pequeno questionário, ou a uma condição de controle (sem espirros ou tossidas). Os resultados indicam que essa manipulação simples influenciou as percepções dos participantes sobre o risco. Os sujeitos na condição com espirros, comparados com a condição sem espirros, avaliaram como mais elevado o seu risco de contrair uma doença séria, seu risco de ter um ataque cardíaco antes dos 50 anos e seu risco de morrer em um crime ou acidente. De maneira interessante, comparados com os sujeitos na condição de controle, os indivíduos na condição com espirros também foram mais prováveis de favorecer gastos federais para vacinas para gripe do que a criação de empregos "verdes". Como esse experimento foi realizado em um ambiente natural, é mais provável que seja representativo das condições do "mundo real". Assim, podemos ter mais confiança de que os resultados podem ser generalizados para outras situações do mundo real do que se fosse criada uma situação artificial no laboratório.

A validade externa dos resultados experimentais também pode ser estabelecida por *replicação parcial*. As replicações parciais costumam ser feitas como uma parte rotineira do processo de investigar as condições em que um fenômeno ocorre. Uma replicação parcial pode ajudar a estabelecer a validade externa, mostrando que um resultado experimental ocorre quando se usam procedimentos experimentais levemente diferentes. Considere o mesmo experimento básico feito em uma grande universidade privada na região metropolitana e uma pequena faculdade comunitária na zona rural; os participantes e situações são muito diferentes. Se forem obtidos os mesmos resultados, mesmo com esses participantes e situações diferentes, podemos dizer que os resultados podem ser generalizados entre essas duas populações e situações. Observe que nenhum dos experimentos tem validade externa por si só; são *os resultados* que ocorrem em *ambos* os experimentos que têm validade externa.

Os pesquisadores também podem estabelecer a validade externa de seus resultados com *replicações conceituais*. O que queremos generalizar a partir de qualquer estudo são relações conceituais entre variáveis, e não as condições, manipulações, situações ou amostras específicas (ver Basgall e Crowder, 1989; Mook, 1983). Anderson e Bushman (1997) apresentaram um exemplo ilustrando a lógica de uma replicação conceitual. Considere um estudo com crianças de 5 anos para determinar se um determinado insulto (fala infantil insultadora: "bobo", "feio") induz raiva e agressividade. Podemos então fazer uma replicação para verificar se o mesmo insulto gera o mesmo resultado com adultos de 35 anos. Conforme afirmam Anderson e Bushman, os resultados para crianças de 5 anos provavelmente não seriam replicados com os adultos de 35 anos porque "fala infantil insultadora simplesmente não tem a mesma 'força' para pessoas de 5 e 35 anos" (p. 21). Todavia, se quisermos estabelecer a validade externa da ideia de que os "insultos aumentam o comportamento agressivo", podemos usar palavras diferentes que sejam insultos significativos para cada população.

Quando Anderson e Bushman (1997) analisaram variáveis relacionadas com a agressividade no nível conceitual, eles observaram que os resultados de experimentos realizados em ambientes laboratoriais e resultados de estudos correlacionais em situações no mundo real eram bastante semelhantes. Os autores concluíram que os experimentos laboratoriais "artificiais" fornecem informações significativas sobre a agressividade, pois demonstram as mesmas relações conceituais que são observadas para a agressividade no mundo real. Além disso, os experimentos laboratoriais permitem que os pesquisadores isolem as causas potenciais da agressividade e investiguem condições limítrofes para quando a agressividade irá ou não ocorrer.

E o que dizer quando os resultados observados no laboratório e no mundo real diferem? Anderson e Bushman (1997) afirmam que essas discrepâncias, em vez de serem evidências da fraqueza de um dos métodos, devem ser usadas para nos ajudar a refinar nossas teorias sobre a agressividade. Ou seja, as discrepâncias devem nos fazer reconhecer que processos psicológicos diferentes podem estar ocorrendo em cada situação. Quando aumentamos a nossa compreensão dessas discrepâncias, aumentamos a nossa compreensão sobre a agressividade.

Seria praticamente impossível estabelecer a validade externa de *cada* estudo em psicologia realizando replicações parciais ou replicações conceituais. Porém, se levarmos a sério argumentos como os de Dawes (1991) e Sue (1999), como realmente deveríamos, parece que estamos enfrentando uma tarefa impossível. Como, por exemplo, podemos mostrar que um resultado experimental obtido com um grupo de estudantes universitários pode ser generalizado para grupos de adultos idosos, profissionais em atividade, indivíduos com menos formação educacional, e assim por diante? Underwood e Shaughnessy (1975) sugerem uma abordagem possível que merece ser considerada. Sua noção é que devemos pressupor que o comportamento seja relativamente contínuo ao longo do tempo, entre sujeitos diferentes e nas várias situações, a menos que tenhamos razão para pensar o contrário. Essencialmente, é mais provável que a validade externa de pesquisas seja estabelecida pelo bom senso da comunidade científica do que por evidências empíricas definitivas.

Desenho de grupos pareados

- Um desenho de grupos pareados pode ser usado para criar grupos comparáveis quando existem poucos sujeitos para que a designação aleatória funcione efetivamente.
- Usar sujeitos pareados em relação à variável dependente é a melhor abordagem para criar grupos pareados, mas o desempenho em qualquer tarefa de pareamento deve estar correlacionado com o teste da variável dependente.

Depois que os sujeitos são combinados no teste de pareamento, eles devem ser designados aleatoriamente às condições da variável independente.

Para funcionar efetivamente, o desenho de grupos aleatórios exige amostras de tamanho suficiente para garantir que as diferenças individuais entre os sujeitos sejam balanceadas pela designação aleatória. Ou seja, a premissa do modelo com grupos aleatórios é que as diferenças individuais se "nivelam" entre os grupos. Mas quantos sujeitos são necessários para que esse processo de nivelamento funcione como deveria? A resposta é "depende". Serão necessários mais sujeitos para nivelar as diferenças individuais quando as amostras forem tiradas de uma população heterogênea do que de uma população homogênea.

Podemos ter relativa confiança de que a designação aleatória *não* será efetiva para equilibrar as diferenças entre sujeitos quando são testados pequenos números de sujeitos de populações heterogêneas. Todavia, essa é exatamente a situação que os pesquisadores enfrentam em várias áreas da psicologia. Por exemplo, alguns psicólogos do desenvolvimento estudam bebês recém-nascidos; outros estudam idosos. Os bebês recém-nascidos e os idosos certamente representam populações diversas, e os psicólogos do desenvolvimento muitas vezes têm números limitados de sujeitos.

Uma alternativa que os pesquisadores têm nessa situação é administrar todas as condições do experimento a todos os sujeitos, usando um desenho com medidas repetidas (a ser discutido no Capítulo 7). Todavia, certas variáveis independentes exigem grupos separados de sujeitos para cada nível. Por exemplo, suponhamos que os pesquisadores desejem comparar dois tipos de cuidado pós-natal para bebês prematuros e não seja possível administrar os dois tipos a cada bebê. Nessa situação, e em muitas outras, os pesquisadores precisam testar grupos separados no experimento.

O **desenho de grupos pareados** é uma boa alternativa quando não é possível usar o desenho de grupos aleatórios e o desenho de medidas repetidas de forma efetiva. A lógica do desenho de grupos pareados é simples e convincente. Em vez de usar designação aleatória para formar grupos comparáveis, o pesquisador torna os grupos equivalentes combinando os sujeitos. Depois que se formaram grupos comparáveis pela combinação, a lógica do desenho com grupos pareados é a mesma que a do desenho de grupos aleatórios (ver Figura 6.4). Na maioria dos usos do desenho de grupos pareados, usa-se uma tarefa pré-teste para combinar os sujeitos. O desafio é selecionar uma tarefa pré-teste (também chamada tarefa de pareamento) que iguale os grupos em uma dimensão que seja relevante para o resultado do experimento. O *desenho de grupos pareados somente tem utilidade quando existe uma boa tarefa de pareamento disponível.*

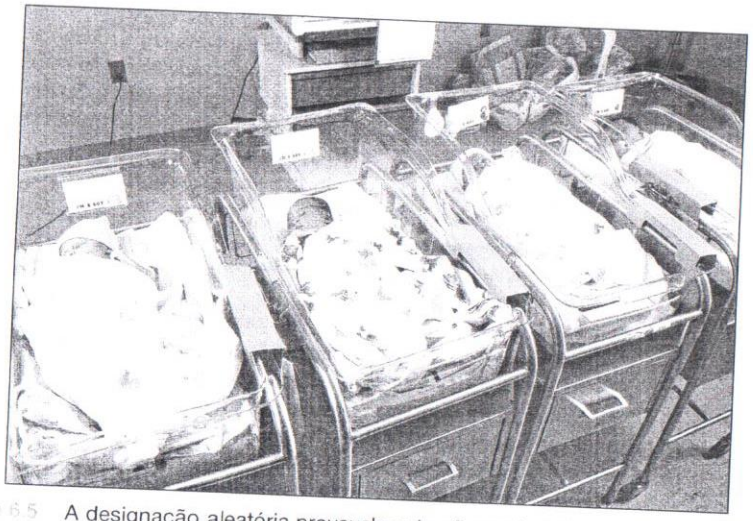

Figura 6.5 A designação aleatória provavelmente não será efetiva para balancear as diferenças entre sujeitos quando são testados pequenos números de sujeitos de populações heterogêneas (p.ex., recém-nascidos). Nessa situação, os pesquisadores talvez devam considerar o uso do desenho de grupos pareados.

A tarefa de pareamento preferida é aquele que usa a mesma tarefa que será usada no experimento propriamente dito. Por exemplo, se a variável dependente do experimento é a pressão sanguínea, os sujeitos devem ser combinados conforme a pressão sanguínea antes do começo do experimento. A combinação é realizada mensurando-se a pressão sanguínea de todos os participantes e depois formando duplas ou trios ou quartetos de participantes (dependendo do número de condições no experimento) com pressão sanguínea idêntica ou muito parecida. Assim, no começo do experimento, os participantes de grupos diferentes têm, *em média*, pressão sanguínea equivalente. Os pesquisadores então podem atribuir ao tratamento quaisquer diferenças grupais na pressão sanguínea observadas no final do estudo (supostamente outras variáveis potenciais foram mantidas constantes ou balanceadas).

Em certos experimentos, não se pode usar a principal variável dependente para combinar os sujeitos. Por exemplo, considere um experimento que ensina aos participantes diferentes abordagens para resolver um quebra-cabeça. Se for usado um pré-teste para ver quanto tempo os indivíduos levam para resolver o jogo, os participantes provavelmente aprenderão a solução durante o pré-teste. Nesse caso, seria impossível observar diferenças na velocidade com que diferentes grupos de participantes resolvem o quebra-cabeça após a manipulação experimental. Nessa situação, a outra melhor alternativa para uma tarefa de pareamento é usar um teste da *mesma classe ou categoria* que o teste experimental. Em nosso experimento com resolução de problemas, os participantes podem ser combinados conforme o seu desempenho ao resolverem um teste diferente do quebra-cabeça experimental. Uma alternativa menos preferida, mas ainda possível, para combinar os sujeitos é usar um teste que seja de uma *classe diferente* do teste experimental. Para nosso experimento com resolução de problemas, os participantes poderiam ser combinados segundo algum teste de capacidade geral, como um teste de capacidade espacial. Todavia, ao usarem essas alternativas, os pesquisadores devem confirmar que o desempenho no teste de pareamento está correlacionado com o desempenho no teste usado como variável dependente. De um modo geral, à medida que diminui a correlação entre o teste de pareamento e a variável dependente, a vantagem do desenho com grupos pareados, em relação ao desenho de grupos aleatórios, também diminui.

Mesmo quando existe um bom teste de pareamento disponível, não é suficiente usar combinação para formar grupos comparáveis em um experimento. Por exemplo, considere um desenho de grupos pareados para comparar dois métodos de tratar de bebês prematuros, de maneira a aumentar seu peso corporal. Seis pares de bebês prematuros podem ser combinados segundo seu peso corporal inicial. Todavia, restam outras características potencialmente relevantes dos participantes, além daquelas medidas pelo teste de pareamento. Por exemplo, os dois grupos de bebês prematuros podem não ser comparáveis em sua saúde geral ou no grau de vínculo parental.

É importante, portanto, usar designação aleatória no desenho de grupos pareados, visando balancear outros fatores potenciais além do teste de pareamento. Especificamente, depois de combinar os bebês segundo o peso corporal, os indivíduos de cada par seriam designados aleatoriamente a um dos dois grupos. Concluindo, *o desenho de grupos pareados é uma alternativa melhor do que o uso de grupos aleatórios quando existe um bom teste de pareamento e quando há somente um pequeno número de sujeitos disponível para um experimento que exija grupos separados para cada condição.*

Desenho de grupos naturais

- Para implementar desenhos de grupos naturais, as variáveis relacionadas com as diferenças individuais (ou variáveis dos sujeitos) são selecionadas, em vez de manipuladas.
- O desenho de grupos naturais representa um tipo de pesquisa correlacional em que os pesquisadores procuram covariações entre variáveis de grupos naturais e variáveis dependentes.
- Não é possível fazer inferências causais relacionadas com os efeitos de variáveis de grupos naturais porque existem explicações alternativas plausíveis para as diferenças grupais.

Os pesquisadores em muitas áreas da psicologia estão interessados em variáveis independentes chamadas de **variáveis de diferenças individuais**, ou *variáveis do sujeito*. Uma variável de diferença individual é uma característica ou traço que varia entre indivíduos. A afiliação religiosa é um exemplo de uma variável de diferença individual. Os pesquisadores não podem manipular essa variável designando pessoas aleatoriamente a grupos católicos, judeus, muçulmanos, protestantes ou outros. Ao contrário, os pesquisadores "controlam" a variável da afiliação religiosa, selecionado sistematicamente indivíduos que pertencem *naturalmente* a esses grupos. As variáveis de diferenças individuais, como gênero, extroversão-introversão, raça ou idade, são variáveis independentes importantes em muitas áreas da psicologia.

É importante diferenciar experimentos que envolvem variáveis independentes cujos níveis são *selecionados* daqueles que envolvem variáveis independentes cujos níveis são *manipulados*. Os experimentos que envolvem variáveis independentes cujos níveis são selecionados – como variáveis relacionadas com diferenças individuais – são chamados de **desenho de grupos naturais**. O desenho de grupos naturais costuma ser usado em situações em que restrições éticas e práticas nos impedem de manipular diretamente as variáveis independentes. Por exemplo, não importa o quanto possamos estar interessados nos efeitos de uma grande cirurgia sobre uma depressão subsequente, não podemos fazer uma grande cirurgia em um grupo designado aleatoriamente de estudantes de introdução à psicologia e depois comparar seus sintomas de depressão com os de outro grupo que não fez a cirurgia! De maneira semelhante, se estivéssemos interessados na relação entre o divórcio e transtornos emocionais, não poderíamos designar pessoas aleatoriamente para se divorciarem. Todavia, usando o desenho de grupos naturais, podemos comparar pessoas que fizeram cirurgia com pessoas que não fizeram. Do mesmo modo, pessoas que decidiram se divorciar podem ser comparadas com pessoas que decidiram permanecer casadas.

Os pesquisadores usam desenhos de grupos naturais para cumprir os dois primeiros objetivos do método científico: descrição e previsão. Por exemplo, estudos mostram que as pessoas que se separam ou divorciam são muito mais prováveis de receber tratamento psiquiátrico do que aquelas que são casadas, viúvas ou que permaneceram solteiras. Com base em estudos como esses, podemos descrever os indivíduos divorciados e casados em termos de transtornos emocionais, e podemos prever qual grupo é mais provável de ter transtornos emocionais.

Podem surgir problemas sérios, contudo, quando os resultados de desenhos de grupos naturais são usados para fazer afirmações causais. Por exemplo, a observação de que as pessoas divorciadas são mais prováveis do que pessoas casadas de receber cuidados psiquiátricos mostra que esses dois fatores covariam. Pode-se considerar que isso significa que o divórcio causa transtornos emocionais. Porém, antes de concluirmos que o divórcio *causa* transtornos emocionais, devemos garantir que foi atendida a condição de ordem temporal para uma inferência causal. Será que o divórcio precede o transtorno emocional, ou o transtorno emocional precede o divórcio? O desenho de grupos naturais não nos diz isso.

O desenho de grupos naturais também traz problemas quando tentamos satisfazer a terceira condição para demonstrar causalidade, eliminar causas alternativas plausíveis. As diferenças individuais estudadas no desenho de grupos naturais geralmente são confundidas – é provável que os grupos de indivíduos difiram em muitas maneiras *além* da variável usada para classificá-los. Por exemplo, os indivíduos que se divorciam e os indivíduos que continuam casados podem diferir com relação a várias características além do seu estado civil, por exemplo, suas práticas religiosas ou circunstâncias financeiras. Qualquer diferença observada entre indivíduos casados e divorciados talvez se deva a outras características, e não ao divórcio. A manipulação feita pela "natureza" raramente é do tipo controlado que esperamos para estabelecer a validade interna de um experimento.

Existem estratégias para fazer inferências causais no desenho de grupos naturais. Uma abordagem efetiva exige que as diferenças individuais sejam estudadas em combinação com variáveis independentes que possam ser manipuladas. Essa combinação de mais de uma variável independente em um experimento exige o uso de um desenho complexo, que descreveremos no Capítulo 8. Por enquanto, reconheça que fazer inferências causais com base no desenho de grupos naturais pode ser traiçoeiro. Embora certos formatos às vezes sejam chamados de "experimentos", existem diferenças importantes entre um experimento envolvendo uma variável de diferenças individuais e um experimento envolvendo uma variável manipulada.

Resumo

Os pesquisadores fazem experimentos para testar hipóteses derivadas de teorias, mas os experimentos também podem ser usados para testar a efetividade de tratamentos ou programas em situações aplicadas. O método experimental é ideal para identificar relações de causa e efeito quando são implementadas adequadamente técnicas de controle de manipulação, manutenção de condições constantes e balanceamento de diferenças.

No Capítulo 6, enfocamos a aplicação dessas técnicas de controle em experimentos em que diferentes grupos de sujeitos recebem tratamentos diferentes representando os níveis da variável independente (ver Figura 6.6). No desenho de grupos aleatórios, os grupos são formados usando procedimentos de randomização, de modo que sejam comparáveis no começo do experimento. Se os grupos apresentam comportamento diferente após a manipulação, e todas as outras condições forem mantidas constantes, presume-se que a variável independente seja responsável pela diferença. A designação aleatória é o método mais comum para formar grupos comparáveis. Distribuindo as características dos sujeitos igualmente entre as condições do experimento, a designação aleatória é uma tentativa de garantir que as diferenças entre os sujeitos sejam balanceadas, ou equilibradas, entre os grupos do experimento. A técnica mais comum para implementar a designação aleatória é a randomização em bloco.

Existem várias ameaças à validade interna de experimentos que envolvem testar grupos independentes. Deve-se evitar testar grupos intactos mesmo quando os grupos

☑ **Figura 6.6** Neste capítulo, apresentamos três desenhos de grupos independentes.

são designados aleatoriamente às condições, pois é provável que o uso de grupos intactos resulte em um fator de confusão. Não se pode permitir que variáveis externas, como diferentes salas ou diferentes experimentadores, confundam a variável independente de interesse.

Uma ameaça mais séria à validade interna do desenho de grupos aleatórios ocorre quando os sujeitos não concluem o experimento. A perda seletiva de sujeitos ocorre quando os sujeitos são perdidos de maneira diferenciada entre as condições, e uma característica do sujeito, relacionada com o resultado do experimento, é responsável pela perda. Podemos ajudar a prevenir essa perda seletiva restringindo os sujeitos àqueles prováveis de concluir o experimento, ou podemos compensar a perda removendo seletivamente alguns sujeitos comparáveis do grupo que não teve perda. As características de demanda e os efeitos do experimentador podem ser minimizados pelo uso dos procedimentos experimentais adequados, mas podem ser controlados com o uso de um controle com placebo e procedimentos duplos-cegos.

A análise de dados e a estatística proporcionam uma alternativa à replicação para determinar se os resultados de um único experimento podem ser usados como evidência para afirmar que uma determinada variável independente teve um efeito sobre o comportamento. A análise de dados envolve o uso de estatísticas descritivas e estatísticas inferenciais. A descrição dos resultados de um experimento geralmente envolve o uso de médias, desvios padrão e medidas do tamanho do efeito. A meta-análise faz uso de medidas do tamanho do efeito para fornecer uma síntese quantitativa dos resultados de um grande número de experimentos sobre uma pergunta de pesquisa importante.

As estatísticas inferenciais são importantes na análise de dados, pois os pesquisadores precisam de um modo de decidir se as diferenças obtidas em um experimento se devem ao acaso ou ao efeito da variável independente. Os intervalos de confiança e o teste da hipótese nula são duas técnicas estatísticas efetivas que os pesquisadores podem usar para analisar experimentos. Todavia, a análise estatística não pode garantir que os resultados experimentais serão significativos ou terão significância prática. A replicação permanece como o teste final da confiabilidade de um resultado de pesquisa.

Os pesquisadores também buscam estabelecer a validade externa de seus resultados experimentais. Ao testarem teorias psicológicas, os pesquisadores tendem a enfatizar a validade interna sobre a validade externa. Uma abordagem efetiva para estabelecer a validade externa dos resultados é selecionar amostras representativas de todas as dimensões que deseja generali-

zar. Por meio de experimentos de campo, os pesquisadores podem aumentar a validade externa de seus estudos para situações do mundo real. As replicações parciais e as replicações conceituais são duas maneiras que os pesquisadores normalmente usam para estabelecer a validade externa.

O desenho de grupos pareados é uma alternativa ao desenho de grupos aleatórios quando existe apenas um pequeno número de sujeitos disponíveis, quando existe uma boa tarefa de pareamento e quando o experimento exige grupos separados para cada tratamento. O maior problema com o desenho de grupos pareados é que os grupos somente são igualados segundo a característica medida pelo teste de pareamento. No desenho de grupos naturais, os pesquisadores selecionam os níveis das variáveis independentes (geralmente diferenças individuais ou variáveis dos sujeitos) e procuram relações sistemáticas entre essas variáveis independentes e outros aspectos do comportamento. Essencialmente, o desenho de grupos naturais envolve procurar correlações entre as características dos sujeitos e seu comportamento. Esses desenhos de pesquisa correlacional trazem problemas no estabelecimento de inferências causais.

Conceitos básicos

validade **interna** 198
desenho **de grupos independentes** 199
designação aleatória 199
desenho **de grupos aleatórios** 199
randomização em bloco 203
ameaças à validade interna 205
perda **mecânica de sujeitos** 207
perda **seletiva de sujeitos** 207
efeitos **do experimentador** 210
grupo **controle com placebo** 210
procedimento **duplo-cego** 211
replicação 212
tamanho do efeito 214
***d* de Cohen** 214
meta-análise 215
teste de significância da hipótese nula 217
estatisticamente significativo 217
intervalo de confiança 218
desenho de grupos pareados 224
variável de diferenças individuais 226
desenho de grupos naturais 226