[capítulo 1]

INTRODUÇÃO AOS TESTES PSICOLÓGICOS E SEUS USOS

O primeiro e mais geral sentido do termo *teste* listado nos dicionários é "exame, observação ou avaliação crítica". Seu sinônimo mais próximo é *prova*. A palavra *crítica*, por sua vez, é definida como "relacionada a... um ponto de virada ou conjuntura especialmente importante" (*Merriam-Webster's collegiate dictionary*, 1995). Não deve nos surpreender, portanto, que quando o termo *psicológico* aparece após a palavra *teste*, a expressão resultante adquira uma conotação um tanto ameaçadora. Os testes psicológicos muitas vezes são utilizados para avaliar indivíduos em algum ponto crítico ou circunstância significativa da vida. Ainda assim, aos olhos de muitas pessoas, os testes parecem ser provações sobre as quais elas pouco sabem e das quais dependem decisões importantes. Em grande parte, este livro visa fornecer aos leitores informações suficientes a respeito dos testes e da testagem psicológica para remover as conotações ameaçadoras e proporcionar meios para que os que fazem uso dos testes psicológicos obtenham mais conhecimento sobre seus usos específicos.

Milhares de instrumentos podem ser chamados corretamente de *testes psicológicos*, mas muitos mais usurpam esta denominação, seja explícita ou sugestivamente. O objetivo básico deste livro é explicar como distinguir os primeiros dos segundos. Por isso, começamos com as características definidoras que os testes psicológicos legítimos de todos os tipos têm em comum. Estas características não apenas definem os testes psicológicos, mas também os diferenciam de outros tipos de instrumentos.

TESTES PSICOLÓGICOS

O *teste psicológico* é um procedimento sistemático para a obtenção de amostras de comportamento relevantes para o funcionamento cognitivo ou afetivo e para a

avaliação destas amostras de acordo com certos padrões. O esclarecimento dos principais termos desta definição é vital para a compreensão de todas as discussões futuras sobre testes. O quadro Consulta Rápida 1.1 explica o sentido e o fundamento lógico de todos os elementos da definição de um teste psicológico. Se alguma das condições mencionadas na definição não for satisfeita, o procedimento em questão não pode ser chamado acuradamente de teste psicológico. No entanto, é importante lembrar que, em essência, os testes psicológicos são simplesmente amostras de comportamento. Tudo o mais se baseia em inferências.

Os testes psicológicos costumam ser descritos como padronizados por dois motivos, ambos contemplam a necessidade de objetividade no processo de testagem. O primeiro está ligado à uniformidade de procedimentos em todos os aspectos importantes da administração, avaliação e interpretação dos testes. Naturalmente, a hora e local em que o teste é administrado, bem como as circunstâncias de sua administração e o examinador que o administra, afetam os resultados. No entanto, o objetivo da padronização dos procedimentos é tornar tão uniformes quanto possíveis todas as variáveis que estão sob o controle do examinador, para que todos que se submetam ao teste o façam da mesma forma.

CONSULTA RÁPIDA 1.1

Elementos básicos da definição de testes psicológicos

Elemento definidor	Explicação	Fundamento
Os testes psicológicos são procedimentos *sistemáticos*.	Caracterizam-se por planejamento, uniformidade e meticulosidade.	Para serem úteis, os testes devem ser objetivos e justos e passíveis de demonstração.
Os testes psicológicos são *amostras de comportamento*.	São pequenos subconjuntos de um todo muito maior.	O uso de amostras de comportamento é eficiente porque o tempo disponível geralmente é limitado.
Os comportamentos avaliados pelos testes são *relevantes para o funcionamento cognitivo, afetivo* ou ambos.	As amostras são selecionadas por sua significância psicológica empírica ou prática.	Os testes, ao contrário dos jogos mentais, existem por sua utilidade; eles são ferramentas.
Os resultados dos testes são *avaliados* e recebem *escores*.	Algum sistema numérico ou categórico é aplicado aos resultados segundo regras preestabelecidas.	Não deve haver dúvidas sobre quais são os resultados de um teste.
Para se avaliar resultados de testes, é necessário ter *padrões* baseados em dados empíricos.	Deve haver uma forma de aplicar um critério ou padrão de comparação comum aos resultados.	Os padrões usados para avaliar os resultados de um teste devem indicar o único sentido dos mesmos.

O segundo sentido da padronização diz respeito ao uso de padrões para a avaliação dos resultados. Estes padrões costumam ser normas derivadas de um grupo de indivíduos – conhecidos como *amostra normativa* ou *de padronização* – no processo de desenvolvimento do teste. O desempenho coletivo do grupo ou grupos de padronização, tanto em termos de médias quanto de variabilidade, é tabulado e passa a ser o padrão pelo qual o desempenho dos outros indivíduos que se submeterem ao teste depois de sua padronização será medido.

Não esqueça

- A palavra *teste* tem múltiplos sentidos.
- O termo *teste psicológico* tem um sentido muito específico.
- Neste livro, a palavra *teste* será usada para se referir a todos os instrumentos que se encaixam na definição de *teste psicológico*.
- Os testes que avaliam habilidades, conhecimentos ou qualquer outra função cognitiva serão referidos como *testes de habilidade*, e todos os outros serão denominados *testes de personalidade*.

Estritamente falando, o termo *teste* deveria ser usado apenas para aqueles procedimentos nos quais as respostas do testando são avaliadas tendo por base sua correção ou qualidade. Tais instrumentos sempre envolvem a avaliação de algum aspecto do funcionamento cognitivo, conhecimento, habilidades ou capacidades de uma pessoa. Por outro lado, instrumentos cujas respostas não são avaliadas como certas ou erradas e cujos testandos não recebem escores de aprovação ou reprovação são denominados *inventários*, *questionários*, *levantamentos*, *listas de verificação*, *esquemas* ou *técnicas projetivas*, e geralmente são agrupados sob a rubrica de *testes de personalidade*. Estes são ferramentas delineadas para se obter informações a respeito das motivações, preferências, atitudes, interesses, opiniões, constituição emocional e reações características de uma pessoa a outras pessoas, situações ou estímulos. Tipicamente, são compostos de perguntas de múltipla escolha ou verdadeiro-falso, exceto as técnicas projetivas, que usam perguntas abertas. Também podem envolver escolhas forçadas entre afirmações que representam alternativas contrastantes, ou a determinação do grau em que o testando concorda ou discorda com várias afirmações. Na maior parte das vezes, os inventários de personalidade, questionários e outros instrumentos do gênero são de auto-relato, mas alguns também são delineados de modo a eliciar relatos de outros indivíduos que não da pessoa que está sendo avaliada (por exemplo, um dos pais, o cônjuge ou professor). Por conveniência, e de acordo com o uso comum, neste livro o termo *teste* vai ser aplicado a todos os instrumentos, independentemente do tipo, que se encaixem na definição de teste psicológico. Testes que avaliam conhecimentos, habilidades ou funções cognitivas serão designados como *testes de habilidades*, e todos os outros serão referidos como *testes de personalidade*.

Outros termos usados em relação a testes e títulos de testes

Alguns outros termos que são usados em relação a testes, às vezes de forma pouco precisa, justificam uma explicação. Um deles é a palavra *escala*, que pode se referir a:

- um teste composto de várias partes, como, por exemplo, a *Escala de Inteligência Stanford-Binet*;
- um subteste, ou conjunto de itens dentro de um teste, que mede uma característica distinta e específica, como, por exemplo, a *Escala de Depressão* do Inventário Multifásico Minnesota de Personalidade (MMPI);
- um conjunto de subtestes que compartilham certas características, como, por exemplo, as *Escalas Verbais* dos testes de inteligência Wechsler;
- um instrumento separado formado por itens que avaliam uma única característica, como, por exemplo, a *Escala de Lócus de Controle Interno-Externo* (Rotter, 1966) ou
- sistema usado para classificar ou atribuir valor a alguma dimensão mensurável, como, por exemplo, uma *escala* de 1 a 5 na qual 1 significa *discordo totalmente*, e 5 significa *concordo totalmente*.

Como se vê, quando usado em referência a testes psicológicos, o termo *escala* tornou-se ambíguo e pouco preciso. No entanto, no campo da mensuração psicológica – também conhecido como *psicometria* – *escala* tem um sentido mais preciso. Refere-se a um grupo de itens que diz respeito a uma única variável e são dispostos em ordem de dificuldade ou intensidade. O processo de se chegar ao seqüenciamento dos itens é denominado *escalonamento* (*scaling*).

Bateria é outro termo usado com freqüência nos títulos de testes. Uma bateria é um grupo de vários testes ou subtestes que são aplicados de uma única vez a uma única pessoa. Quando diversos testes são reunidos por uma editora para serem utilizados para um fim específico, a palavra *bateria* geralmente aparece no título, e o grupo de testes é visto como um único instrumento. Diversos exemplos deste uso ocorrem em instrumentos neuropsicológicos (como a *Bateria Neuropsicológica Halstead-Reitan*), nos quais muitas funções cognitivas precisam ser avaliadas por meio de testes separados para detectar um possível comprometimento cerebral. O termo *bateria* também é usado para designar qualquer grupo de testes individuais selecionados especificamente por um psicólogo, para uso com um determinado cliente, na tentativa de responder à questão específica que gerou seu encaminhamento, geralmente de natureza diagnóstica.

OS TESTES PSICOLÓGICOS COMO FERRAMENTAS

O fato mais básico a respeito dos testes psicológicos é que eles são ferramentas. Isso significa que sempre são um meio para alcançar um fim, e nunca um fim em si mesmos. Como outras ferramentas, os testes psicológicos podem ser extremamente úteis – e até mesmo insubstituíveis – quando usados de forma apropriada e hábil.

No entanto, também podem ser mal-aplicados, podendo limitar ou anular sua utilidade e, por vezes, até mesmo resultam em conseqüências prejudiciais.

Uma boa maneira de ilustrar as semelhanças entre os testes e outras ferramentas mais simples é a analogia entre um teste e um martelo. Ambos são ferramentas para fins específicos, mas podem ser usados de várias formas. O martelo é útil basicamente para fixar pregos em superfícies variadas. Quando usado corretamente para seu fim específico, o martelo pode ajudar a construir uma casa, montar um móvel, pendurar quadros em uma galeria e fazer muitas outras coisas. Os testes psicológicos são ferramentas criadas para ajudar na obtenção de inferências a respeito de indivíduos ou grupos, e, quando usados corretamente, podem ser componentes-chave na prática e na ciência da psicologia.

Assim como os martelos podem ser usados para fins positivos diferentes daqueles para os quais foram criados (por exemplo, como pesos de papel ou calços de porta), os testes psicológicos também podem servir a outros fins além daqueles para os quais foram originalmente criados, como aumentar o autoconhecimento e a autocompreensão. Além disso, assim como os martelos podem machucar pessoas e destruir coisas, quando usados com incompetência ou maldade, os testes psicológicos também podem ser usados de formas danosas. Quando seus resultados são mal-interpretados ou mal-utilizados, podem prejudicar pessoas, rotulando-as de maneira injustificada, negando-lhes oportunidades injustamente ou simplesmente desencorajando-as.

Todas as ferramentas, sejam martelos ou testes, podem ser avaliadas segundo a qualidade de seu delineamento e construção. Quando examinados deste ponto de vista, antes de serem usados, os testes são avaliados somente em um sentido técnico limitado, e sua avaliação é de interesse principalmente para os usuários em potencial. Depois que são colocados em uso, no entanto, os testes não podem mais ser avaliados independentemente das habilidades de seus usuários, do modo como são usados e dos fins a que servem. Esta avaliação durante o uso muitas vezes envolve questões de direcionamento, valores sociais e até mesmo prioridades políticas. É neste contexto que a avaliação do uso de testes adquire significado prático para uma gama mais ampla de públicos.

Não esqueça

Os testes psicológicos são avaliados em *dois* momentos distintos e de *duas* formas diferentes:

1. Quando estão sendo considerados como ferramentas em potencial por possíveis usuários. Neste momento, suas qualidades técnicas são a principal preocupação.
2. Depois que são colocados em uso para um objetivo específico. Neste momento, a habilidade do usuário e o modo como os testes são usados são as principais considerações.

Padrões de testagem

Devido à importância singular dos testes para todos os profissionais que os utilizam e para o público em geral desde meados dos anos de 1950, três importantes

organizações profissionais uniram forças para promulgar padrões que ofereçam uma base para a avaliação de testes, práticas de testagem e efeitos de seu uso. A versão mais recente são os *Padrões de testagem educacional e psicológica* (*Standards for Educational and Psychological Testing*), publicados em 1999 pela Associação Americana de Pesquisa em Educação (AERA [*American Educational Research Association*]) e preparados conjuntamente pela AERA e pela Associação Americana de Psicologia (APA [*American Psychological Association*]) e pelo Conselho Nacional de Mensurações em Educação (NCME [*National Council on Measurement in Education*]). Como indica o quadro Consulta Rápida 1.2, esses padrões são citados ao longo de todo este livro e serão referidos daqui por diante como *Padrões de Testagem*.

CONSULTA RÁPIDA 1.2 Padrões de Testagem

- Esta designação será usada freqüentemente ao longo deste livro em referência aos *Padrões de testagem educacional e psicológica*, publicados conjuntamente em 1999 pela Associação Americana de Pesquisa em Educação, Associação Americana de Psicologia e Conselho Nacional de Mensurações em Educação.
- Os *Padrões de Testagem* são a fonte mais importante de critérios para a avaliação de testes, práticas de testagem e efeitos do uso de testes.

OS TESTES PSICOLÓGICOS COMO PRODUTOS

O segundo fato mais básico a respeito dos testes psicológicos é que eles são produtos, mas a maioria das pessoas não atentam para isso. Os testes são produtos comercializados e usados primariamente por psicólogos e educadores profissionais, assim como as ferramentas da odontologia são comercializadas para dentistas. O público leigo não está ciente da natureza comercial dos testes psicológicos porque eles são anunciados em publicações e catálogos voltados para as categorias profissionais que os utilizam. Não obstante, permanece o fato de que muitos, senão a maioria dos testes psicológicos são concebidos, desenvolvidos, anunciados e vendidos para fins aplicados no contexto da educação, administração ou saúde mental, e devem gerar lucros para as pessoas que os produzem como qualquer outro produto comercial.

Como veremos, desde o início o mercado dos testes psicológicos foi impulsionado principalmente pela necessidade de se tomar decisões práticas a respeito de pessoas. Uma vez que os testes são ferramentas profissionais que podem ser usadas para beneficiar pessoas e como produtos comerciais, alguns esclarecimentos sobre as várias partes envolvidas no negócio da testagem e seus papéis são justificados. O quadro Consulta Rápida 1.3 mostra uma lista dos principais participantes do processo de testagem e seus respectivos papéis.

CONSULTA RÁPIDA 1.3

Os participantes do processo de testagem e seus papéis

Participantes	Seus papéis no processo de testagem
Autores e criadores	Concebem, preparam e criam os testes. Também encontram formas de divulgar seu trabalho publicando os testes comercialmente ou por meio de publicações profissionais como livros e periódicos.
Editoras	Publicam, anunciam e vendem os testes, controlando sua distribuição.
Revisores	Preparam críticas e avaliações dos testes baseados em seus méritos técnicos e práticos.
Usuários	Selecionam ou decidem usar um teste específico para algum objetivo. Também podem desempenhar outros papéis, como, por exemplo, examinadores ou avaliadores.
Examinadores ou administradores	Administram o teste individualmente ou em grupo.
Testandos	Submetem-se ao teste por escolha ou necessidade.
Avaliadores	Computam as respostas do testando e as transformam em escores de teste por meios objetivos ou mecânicos ou aplicação de julgamentos de avaliação.
Intérpretes de resultados	Interpretam os resultados dos testes para seus consumidores finais, que podem ser testandos individuais ou seus parentes, outros profissionais ou organizações de vários tipos.

Conforme estipulam os *Padrões de Testagem*, "os interesses das várias partes envolvidas no processo de testagem geralmente são congruentes, mas não sempre" (AERA, APA, NCME, 1999, p.1). Por exemplo, os *autores* dos testes geralmente são acadêmicos ou pesquisadores interessados na teorização ou pesquisa psicológica, mais do que nas aplicações práticas ou lucros. Os *usuários* dos testes estão mais interessados na adequação e utilidade dos materiais que usam para seus próprios fins, enquanto que as *editoras* estão naturalmente inclinadas a considerar mais importante o lucro a ser obtido com a venda dos testes. Além disso, os participantes do processo de testagem podem desempenhar um ou mais dos vários papéis descritos no quadro Consulta Rápida 1.3. Os usuários de testes podem administrar, avaliar e interpretar os resultados dos testes que selecionaram ou podem delegar uma ou mais dessas funções a outros sob sua supervisão. Da mesma forma, as editoras podem contratar criadores de testes que criem instrumentos para os quais acreditam que exista mercado. Não obstante, de todos os participantes do processo de testagem, os *Padrões de Testagem* atribuem "a responsabilidade última pelo uso e interpretação apropriados dos testes" predominantemente ao seu usuário (p.112).

HISTÓRICO DA TESTAGEM PSICOLÓGICA

Muito embora os testes psicológicos possam ser usados para explorar e investigar uma ampla gama de variáveis, seu uso mais básico e típico é como ferramenta na tomada de decisões que envolvem pessoas. Não é coincidência que os testes psicológicos como os conhecemos hoje tenham surgido no início do século XX. Antes do estabelecimento de sociedades urbanas, industriais e democráticas havia pouca necessidade de que a maioria das pessoas tomasse decisões a respeito de outras, além daquelas de sua família imediata ou do círculo próximo de conhecidos. Nas sociedades rurais, agrárias e autocráticas, as principais decisões sobre a vida dos indivíduos eram tomadas em grande parte por pais, mentores, governantes e, acima de tudo, segundo o gênero, a classe, o local e as circunstâncias nas quais as pessoas nasciam. Mesmo assim, muito antes do século XX já existiam diversos precursores interessantes da moderna testagem psicológica em várias culturas e contextos.

Antecedentes da testagem moderna no campo ocupacional

Um problema perene em qualquer campo de trabalho é a questão de como selecionar os melhores candidatos possíveis para um determinado emprego. Os precursores mais antigos da testagem psicológica são encontrados precisamente nesta área, no sistema de exames competitivos desenvolvido pelo antigo império chinês para selecionar indivíduos merecedores de posições governamentais. Este precursor dos modernos procedimentos de seleção de pessoal remonta aproximadamente ao ano 200 a.C. e passou por diversas transformações ao longo de sua história (Bowman, 1989). Os concursos para o serviço público chinês envolviam demonstrações de proficiência em música, uso do arco e habilidades de montaria, entre outras coisas, bem como exames escritos sobre temas como leis, agricultura e geografia. Aparentemente, o ímpeto para o desenvolvimento deste sistema sofisticado de utilização de recursos humanos – aberto a qualquer indivíduo que fosse recomendado ao imperador por autoridades locais de todo o império – era o fato de que a China não tinha o tipo de classe dominante hereditária que era comum na Europa até o século XX. O sistema de concursos da China imperial foi abolido em 1905 e substituído por seleções baseadas em estudos universitários, mas serviu como inspiração para as seleções ao serviço público desenvolvidas na Inglaterra na década de 1850, as quais, por sua vez, estimularam a criação do Concurso Nacional para o Serviço Público dos Estados Unidos (*U.S. Civil Service Examination*) na década de 1860 (DuBois,1970).

Antecedentes da testagem moderna no campo da educação

Uma das questões mais básicas em qualquer contexto educacional é como determinar se os estudantes adquiriram o conhecimento ou as habilidades que seus professores tentaram lhes transmitir. Por isso, não surpreende que o primeiro uso da testagem no campo da educação tenha ocorrido na Idade Média, com o surgimento

das primeiras universidades européias no século XIII. Por volta desta época, o grau universitário passou a ser usado como forma de certificar a qualificação para ensinar, e exames orais formais começaram a ser aplicados para dar aos candidatos a um diploma a oportunidade de demonstrar sua competência (DuBois, 1970). Pouco a pouco, o uso de exames se disseminou para o nível médio da educação e, à medida que o papel se tornou mais barato e disponível, provas escritas substituíram os exames orais na maioria dos contextos educacionais. Ao final do século XIX, tanto na Europa quanto nos Estados Unidos, estas provas já estavam bem-estabelecidas como método para determinar quem deveria receber um diploma universitário e quem estava capacitado a exercer profissões liberais como medicina ou direito.

Antecedentes da testagem moderna na psicologia clínica

Outra questão humana fundamental que pôde ser e foi contemplada com a testagem psicológica constituiu-se no problema de diferenciar o "normal" do "anormal" nas áreas intelectual, emocional e comportamental. No entanto, em contraste com o contexto ocupacional ou educacional, em que as bases sobre as quais as decisões são tomadas tradicionalmente foram bastante claras, o campo da psicopatologia continuou envolto em mistério e misticismo por um período muito mais longo.

Diversos antecedentes dos testes psicológicos se originaram no campo da psiquiatria (Bondy, 1974). Muitos destes primeiros testes foram desenvolvidos na Alemanha na segunda metade do século XIX, embora alguns deles datem do início daquele século e tenham sido elaborados na França. Quase invariavelmente, esses instrumentos foram criados com o fim expresso de avaliar o nível de funcionamento cognitivo de pacientes com vários tipos de transtornos, como retardo mental ou danos cerebrais. Entre as amostras de comportamento usadas nestes primeiros testes havia questões a respeito do sentido de provérbios e diferenças e semelhanças entre pares de palavras, bem como tarefas de memória como a repetição de séries de dígitos apresentadas oralmente. Muitas técnicas desenvolvidas no século XIX eram engenhosas, tendo sobrevivido e sido incorporadas em testes modernos que ainda estão em uso (ver McReynolds, 1986).

A despeito de sua engenhosidade, os criadores dos primeiros precursores dos testes clínicos eram prejudicados por pelo menos dois fatores. Um deles era a escassez de conhecimentos – e a abundância de superstições e concepções equivocadas – relacionada à psicopatologia. Neste aspecto, por exemplo, a distinção entre psicose e retardo mental não foi formulada claramente até 1838, quando o psiquiatra francês Esquirol sugeriu que a capacidade de usar a linguagem é o critério mais confiável para estabelecer o nível de funcionamento mental de uma pessoa. Um segundo fator que impediu a disseminação ampla do uso dos primeiros testes psiquiátricos era sua falta de padronização em termos de procedimentos e de um referencial uniforme com o qual interpretar os resultados. Em grande parte, as técnicas desenvolvidas por neurologistas e psiquiatras do século XIX como Guislain, Snell, von Grashey, Rieger e outros foram criadas com o objetivo de examinar um paciente ou uma população de pacientes específicos. As amostras de comporta-

mento eram coletadas de forma não-sistemática e interpretadas pelos clínicos a partir do seu julgamento profissional, e não em referência a dados normativos (Bondy, 1974).

Um avanço significativo foi obtido na psiquiatria durante a década de 1890, quando Emil Kraepelin dedicou-se a classificar os transtornos mentais segundo suas causas, sintomas e cursos. Kraepelin queria aplicar o método científico à psiquiatria, e sua contribuição foi vital para delinear os quadros clínicos da esquizofrenia e do transtorno bipolar, os quais, na época, eram conhecidos respectivamente como *dementia praecox* e *psicose maníaco-depressiva*. Ele propôs um sistema para comparar indivíduos sãos e afetados a partir de características como distratibilidade, sensibilidade e capacidade de memória, e foi o pioneiro no uso da técnica da associação livre com pacientes psiquiátricos. Embora alguns alunos de Kraepelin tenham criado uma bateria de testes e continuado a se dedicar aos objetivos do mestre, os resultados de seu trabalho não foram tão frutíferos quanto eles esperavam (DuBois, 1970).

Antecedentes da testagem moderna na psicologia científica

As investigações dos psicofísicos alemães Weber e Fechner em meados do século XIX iniciaram uma série de avanços que culminaram na criação, por Wilhelm Wundt em 1879, do primeiro laboratório dedicado à pesquisa de natureza puramente psicológica, em Leipzig. Este evento é considerado por muitos o início da psicologia como disciplina formal, distinta da filosofia. Com o surgimento da nova disciplina da psicologia experimental, também surgiu muito interesse no desenvolvimento de aparatos e procedimentos padronizados para mapear a gama das capacidades humanas no campo da sensação e da percepção. Os primeiros psicólogos experimentais estavam interessados em descobrir as leis gerais que governavam as relações entre os mundos físico e psicológico. Eles tinham pouco ou nenhum interesse nas diferenças individuais – o principal item de interesse na psicologia diferencial e na testagem psicológica – as quais, na verdade, tendiam a ver como fonte de erros. Não obstante, sua ênfase na necessidade de precisão das mensurações e de condições padronizadas de laboratório provaria ser uma contribuição importante para o campo incipiente da testagem psicológica.

O laboratório de Wundt na Alemanha floresceu nas últimas décadas do século XIX e treinou muitos psicólogos dos Estados Unidos e de outros países, que voltaram para seus locais de origem para estabelecer laboratórios semelhantes. Por volta da mesma época, um inglês chamado Francis Galton interessou-se pela mensuração das funções psicológicas a partir de uma perspectiva inteiramente diferente. Galton era um homem de grande curiosidade intelectual e muitas realizações, cuja posição social e financeira privilegiada lhe permitia dedicar-se a uma ampla gama de interesses. Também era primo e grande admirador de Charles Darwin, cuja teoria da evolução das espécies por meio da seleção natural tinha revolucionado as ciências da vida em meados do século XIX. Após ler o tratado de seu primo sobre a origem das espécies, Galton decidiu investigar a noção de que os dons intelectuais tendem a se transmitir de uma geração à outra. Para este fim, montou um laboratório

rio antropométrico em Londres, no qual por vários anos coletou dados sobre uma série de características físicas e psicológicas – como envergadura dos braços, altura, peso, capacidade vital, força e acuidade sensorial de vários tipos – de milhares de indivíduos e famílias. Galton estava convencido de que a capacidade intelectual é uma função da agudeza de sentidos de cada pessoa para perceber e discriminar estímulos, que, por sua vez, seria de natureza hereditária. Por meio da acumulação e tabulação cruzada de seus dados antropométricos, Galton pretendia estabelecer tanto a gama de variação destas características como suas inter-relações e concordância entre indivíduos com diferentes graus de laços familiares (Fancher, 1996).

Galton não teve sucesso em seu objetivo final, que era promover a *eugenia*, um campo de estudos que ele criara com o objetivo de melhorar a raça humana por meio da reprodução seletiva de seus espécimes mais aptos. Com este fim em mente, ele queria descobrir uma forma de avaliar a capacidade intelectual de crianças e adolescentes através de testes, para identificar desde cedo os indivíduos melhor dotados e encorajá-los a gerar muitos filhos. Mesmo assim, o trabalho de Galton foi continuado e consideravelmente ampliado nos Estados Unidos por James McKeen Cattell, que também tentou sem sucesso ligar várias medidas de poder discriminativo, perceptivo e associativo (que ele denominava testes "mentais") a estimativas independentes de nível intelectual, como notas escolares.

À luz de alguns eventos do século XX, como os ocorridos na Alemanha nazista, os propósitos de Galton parecem ser moralmente ofensivos para a maioria das sensibilidades contemporâneas. No entanto, na época em que ele cunhou o termo *eugenia* e anunciou seus objetivos, o potencial genocida de sua iniciativa não foi percebido de modo geral, e muitos indivíduos ilustres do período tornaram-se eugenistas entusiasmados. No processo de seus estudos, por mais equivocados que nos pareçam hoje, Galton conseguiu fazer contribuições significativas para o campo da estatística e da mensuração psicológica. Ao tabular dados comparando pais e filhos, por exemplo, ele descobriu os fenômenos da regressão e da correlação, que forneceram a base para muitas pesquisas psicológicas e análises de dados posteriores. Ele também inventou dispositivos para a mensuração da acuidade auditiva e discriminação de peso, e foi pioneiro no uso de questionários e da associação de palavras na pesquisa psicológica. Como se estas realizações não fossem suficientes, Galton também foi o primeiro a utilizar o método de estudo com gêmeos que, depois de aperfeiçoado, viria a se tornar uma ferramenta de pesquisa primária em genética comportamental.

Uma contribuição adicional ao campo nascente da testagem psicológica ao final do século XIX merece menção especial por que viria a conduzir diretamente ao primeiro instrumento bem-sucedido da moderna era da testagem. Enquanto estudava os efeitos da fadiga na capacidade mental das crianças, o psicólogo alemão Hermann Ebbinghaus – mais conhecido por suas pesquisas inovadoras no campo da memória – elaborou uma técnica conhecida como Teste de Complementação de Ebbinghaus, no qual as crianças deviam preencher lacunas em passagens de textos de onde palavras ou fragmentos de palavras haviam sido omitidos. A importância deste método, que mais tarde seria adaptado para uma variedade de diferentes propósitos, é dupla. Primeiro, como era aplicado a classes inteiras de crianças simultaneamente, o instrumento prenunciou o desenvolvimento dos testes em gru-

po. O mais importante, no entanto, é que provou ser um termômetro eficiente da capacidade intelectual, pois os escores dele derivados correspondiam bem à capacidade mental dos alunos determinada por suas notas em aula. Como resultado disto, Alfred Binet foi inspirado a usar a técnica do completamento e outras tarefas mentais complexas para desenvolver a escala que se tornaria o primeiro teste de inteligência bem-sucedido (DuBois, 1970).

O surgimento da moderna testagem psicológica

No início do século XX, todos os elementos necessários para o surgimento dos primeiros testes psicológicos verdadeiramente modernos e bem-sucedidos estavam presentes:

- Os testes laboratoriais e ferramentas geradas pelos primeiros psicólogos experimentais na Alemanha.
- Os instrumentos de mensuração e técnicas estatísticas desenvolvidos por Galton e seus alunos para coleta e análise de dados sobre diferenças individuais.
- A acumulação de achados significativos nas ciências da psicologia, psiquiatria e neurologia.

Todos estes avanços proporcionaram as bases para o surgimento da testagem moderna, mas seu ímpeto veio da necessidade prática de se tomar decisões de cunho educacional.

Em 1904, o psicólogo francês Alfred Binet foi indicado para uma comissão encarregada de criar um método para avaliar crianças que, devido ao retardo mental ou outros atrasos no desenvolvimento, não conseguiam se beneficiar das classes regulares do sistema educacional público e necessitavam de educação especial. Binet estava particularmente bem preparado para esta tarefa, pois vinha há muito investigando as diferenças individuais por meio de uma variedade de mensurações físicas e fisiológicas, bem como por testes de processos mentais mais complexos, como memória e compreensão verbal. Em 1905, Binet e seu colaborador Theodore Simon publicaram o primeiro instrumento útil para a mensuração da capacidade cognitiva geral, ou inteligência global. A escala Binet-Simon de 1905, como passou a ser conhecida, era uma série de 30 testes ou tarefas de conteúdo e dificuldade variados com o objetivo principal de avaliar o julgamento e a capacidade de raciocínio independentemente da aprendizagem escolar. Incluía perguntas ligadas a vocabulário, compreensão, diferenças entre pares de conceitos, etc., bem como tarefas que englobavam repetir séries de números, seguir instruções, completar passagens fragmentadas de texto e desenhar.

A escala Binet-Simon teve sucesso porque combinava características dos primeiros instrumentos de uma forma nova e sistemática, sendo mais abrangente do que os anteriores, dedicados à avaliação de habilidades mais limitadas. Na verdade, era uma pequena bateria de testes cuidadosamente selecionados, dispostos em ordem de dificuldade e acompanhados por instruções precisas para sua adminis-

tração e interpretação. Binet e Simon administraram a escala a 50 crianças normais entre 3 e 11 anos, bem como a crianças com vários graus de retardo mental. Os resultados destes estudos provaram que a dupla havia criado um procedimento de amostragem do funcionamento cognitivo através do qual o nível geral de capacidade intelectual de uma criança poderia ser descrito quantitativamente, em termos da faixa etária à qual seu desempenho na escala correspondia. A necessidade de tal ferramenta era tão aguda que a escala de 1905 foi rapidamente traduzida para outras línguas e adaptada para uso fora da França.

O nascimento do QI

O próprio Binet revisou, ampliou e aperfeiçoou sua primeira escala em 1908 e 1911. O cálculo do escore evoluiu para um sistema no qual o crédito referente aos itens corretos era apresentado em termos de anos e meses, de tal modo que o *nível mental* atingido representasse a qualidade do desempenho. Em 1911, um psicólogo alemão chamado William Stern propôs que o nível mental obtido na escala Binet-Simon, rebatizado de *escore de idade mental*, fosse dividido pela idade cronológica do sujeito para se obter um quociente mental que representaria de forma mais precisa a capacidade em diferentes idades. Para eliminar a casa decimal, o quociente mental era multiplicado por 100, e logo se tornou conhecido como *quociente de inteligência*, ou *QI*. Este escore agora tão familiar, o *QI-razão*, foi popularizado através do seu uso na revisão mais famosa das escalas Binet-Simon – a Escala de Inteligência Stanford-Binet – publicada em 1916 por Lewis Terman. Apesar dos diversos problemas do QI-razão, seu uso iria continuar por várias décadas, até que uma forma melhor de integrar a idade na pontuação dos testes de inteligência (descrita no Capítulo 3) fosse desenvolvida por David Wechsler (Kaufman, 2000; Wechsler, 1939). A idéia básica de Binet – qual seja, que estar na média, abaixo da média ou acima da média em termos de inteligência significa que um indivíduo tem um desempenho acima, abaixo ou correspondente ao nível típico de sua faixa etária nos testes de inteligência – sobreviveu e tornou-se uma das formas primárias de avaliação da inteligência.

Enquanto Binet desenvolvia suas escalas na França, na Inglaterra, Charles Spearman (ex-aluno de Wundt e seguidor de Galton) vinha tentando provar empiricamente a hipótese de Galton a respeito da ligação entre inteligência e acuidade sensorial. Neste processo, ele havia desenvolvido e ampliado o uso dos métodos de correlação propostos por Galton e Karl Pearson e elaborado as bases conceituais da *análise fatorial*, uma técnica para reduzir um grande número de variáveis a um conjunto menor de fatores que se tornaria central para o avanço da testagem e da teoria dos traços.

Spearman também criou uma teoria da inteligência que enfatizava um fator geral de inteligência (ou *g*) presente em todas as atividades intelectuais (Spearman, 1904a, 1904b). Ele havia obtido um respaldo moderado para as idéias de Galton ao correlacionar as notas e avaliações feitas por professores com medidas de acuidade sensorial, mas logo percebeu que as tarefas propostas na escala Binet-Simon ofereciam uma forma muito mais útil e confiável de avaliar a inteligência

do que as ferramentas que vinha usando. Muito embora Spearman e Binet diferissem muito na forma de ver a natureza da inteligência, suas contribuições combinadas são insuperáveis como motores do desenvolvimento da testagem psicológica no século XX.

A testagem em grupo

Quando Binet morreu em 1911, já tinha considerado a possibilidade de adaptar sua escala a outros usos e de desenvolver testes que pudessem ser administrados por um único examinador a grupos grandes, para uso nas Forças Armadas e outros contextos. A concretização desta idéia, no entanto, não aconteceria na França, mas nos Estados Unidos, onde a escala Binet-Simon havia sido rapidamente traduzida e revisada para uso primordial com crianças em idade escolar, para os mesmos objetivos para os quais fora desenvolvida na França.

Com a entrada dos Estados Unidos na Primeira Guerra Mundial em 1917, o presidente da APA, Robert Yerkes, organizou uma comissão de psicólogos para ajudar no esforço de guerra. Foi decidido que a contribuição mais prática seria desenvolver um teste grupal de inteligência que pudesse ser eficientemente administrado a todos os recrutas do exército dos Estados Unidos, para ajudar na alocação de pessoal. A comissão, formada pelos principais especialistas em testes da época, incluindo Lewis Terman, apressadamente montou e testou um instrumento que veio a ser conhecido como *Army Alpha*, que consistia em oito subtestes que mediam capacidades verbais, numéricas e de raciocínio, bem como julgamento prático e informações gerais. O teste, que seria administrado a mais de um milhão de recrutas, fazia uso de materiais de vários outros instrumentos, incluindo as escalas Binet. Para construí-lo, a comissão se baseou principalmente em um protótipo de teste grupal inédito desenvolvido por Arthur Otis, que tinha criado itens de múltipla escolha que podiam ser pontuados objetiva e rapidamente.

O *Army Alpha* provou-se extremamente útil, e foi seguido rapidamente pelo *Army Beta*, um teste supostamente equivalente, mas que não demandava leitura, e por isso podia ser usado com recrutas analfabetos ou que não falassem inglês. Infelizmente, a pressa com que esses testes foram desenvolvidos e colocados em uso resultou em uma série de práticas de testagem impróprias. Além disso, conclusões injustificadas foram feitas a partir de quantidades maciças de dados que se acumularam rapidamente (Fancher, 1985). Algumas conseqüências negativas do modo como o programa de testagem militar e outros esforços de testagem em massa daquela época foram implementados prejudicaram quase irreversivelmente a reputação da testagem psicológica. Mesmo assim, por meio dos erros cometidos no início da história da testagem moderna, houve um grande aprendizado que mais tarde serviu para a correção e o aperfeiçoamento das práticas na área. Além disso, com os testes militares, a psicologia saiu decisivamente dos laboratórios e da academia e demonstrou seu enorme potencial de aplicação no mundo real.

Após a Primeira Guerra Mundial, a testagem psicológica se fortaleceu nos Estados Unidos. Otis publicou sua Escala Grupal de Inteligência *(Group Intelligence Scale)*, o teste que tinha servido como modelo para o *Army Alpha* em 1918. E. L.

Thorndike, outro importante pioneiro americano que trabalhava no Teachers College de Columbia, produziu um teste de inteligência para formandos do ensino médio, padronizado com uma amostra mais seleta (calouros de universidade) em 1919. Daí em diante, o número de testes publicados cresceu rapidamente, e os procedimentos de administração e pontuação de testes também foram logo aperfeiçoados. Por exemplo, itens de teste de diferentes tipos começaram a ser apresentados em ordem mista, e não mais em subtestes separados, para que um limite total de tempo pudesse ser determinado, eliminando-se a necessidade da cronometragem de cada subteste. Questões de padronização, como a eliminação de palavras que pudessem ser lidas com diferentes pronúncias em testes de ortografia, passaram a receber atenção, assim como a *confiabilidade* dos testes – um termo que, naquela época, englobava o que atualmente se entende por *fidedignidade* e *validade* (DuBois, 1970).

OUTROS AVANÇOS NA TESTAGEM PSICOLÓGICA

Os êxitos alcançados com os testes militares e as escalas Binet provaram seu valor nos processos de tomada de decisão envolvendo pessoas. Isto rapidamente levou a esforços para a criação de instrumentos para auxiliar em diferentes tipos de decisões. Naturalmente, os locais onde os antecedentes dos testes psicológicos tinham surgido – escolas, clínicas e laboratórios de psicologia – também foram o berço das novas formas e tipos dos modernos testes psicológicos.

Uma revisão completa do histórico da testagem na primeira metade do século XX está além do âmbito deste trabalho. Não obstante, um rápido resumo dos avanços mais importantes é instrutivo tanto por si só quanto para ilustrar a diversidade do campo, mesmo em sua fase inicial.

A testagem padronizada no contexto educacional

À medida que aumentava o número de pessoas desfrutando de oportunidades educacionais em todos os níveis, o mesmo ocorreu com a necessidade de mensurações justas, equânimes e uniformes com as quais avaliar os alunos nos estágios iniciais, intermediários e finais do processo educacional. Os dois principais avanços na testagem educacional padronizada no início do século XX são apresentados nos parágrafos a seguir.

Testes padronizados de realização acadêmica

Elaboradas inicialmente por E.L. Thorndike, estas mensurações vinham sendo desenvolvidas desde a década de 1880, quando Joseph Rice começou a estudar a eficiência do aprendizado nas escolas. A escala de caligrafia de Thorndike, publicada em 1910, inaugurou uma nova modalidade de testagem ao criar uma série de espécimes de caligrafia, variando de muito ruim a excelente, em relação às quais o

desempenho dos sujeitos podia ser comparado. Logo depois viriam testes padronizados com o objetivo de avaliar habilidades de aritmética, leitura e ortografia, até que as mensurações destes e outros aspectos se tornassem banais no ensino fundamental e médio. Hoje em dia, os testes padronizados de realização são usados não apenas no contexto educacional, mas também no licenciamento e certificação ao final da formação profissional e em outras situações, incluindo seleção de pessoal, que requerem a avaliação das capacidades em um dado campo de conhecimento.

Testes de aptidão escolar

Nos anos de 1920, exames objetivos inspirados no teste *Army Alpha* começaram a ser usados, em conjunto com as notas do ensino médio, para fins de admissão em faculdades e universidades. Este importante avanço, que culminou na criação do Teste de Aptidão Escolar (SAT, *School Aptitude Test*) em 1926, prenunciou a chegada de muitos outros instrumentos que são usados para selecionar candidatos para cursos de pós-graduação e formação profissional. Entre os exemplos mais conhecidos desse tipo de teste estão o *Graduate Record Exam* (GRE), o *Medical College Admission Test* (MCAT) e o *Law School Admission Test* (LSAT), usados por programas de doutorado, escolas médicas e escolas de direito, respectivamente. Embora todos estes testes contenham partes dedicadas ao seu campo de estudos específico, eles têm em comum a ênfase nas habilidades verbais, quantitativas e de raciocínio necessárias para o sucesso na maioria das iniciativas acadêmicas. É interessante notar que, embora seu objetivo seja diferente daquele dos testes padronizados de realização, seu conteúdo muitas vezes é semelhante. As informações contidas em Consulta Rápida 1.4 apresentam um relato fascinante do histórico da testagem para admissão na educação superior nos Estados Unidos.

CONSULTA RÁPIDA 1.4

O grande teste

O livro de Nicholas Lemann *O grande teste: A história secreta da meritocracia americana* (1999) usa os programas de testagem para admissão universitária, especialmente o SAT, para ilustrar as conseqüências intencionais ou não que estes programas podem ter para a sociedade. O uso em larga escala de escores de testes padronizados para decidir sobre as admissões nas principais instituições de ensino superior foi proposto pela primeira vez por James Bryant Conant, reitor da Universidade de Harvard, e Henry Chauncey, primeiro presidente do Serviço de Testagem Educacional (ETS, *Educational Testing Service*), nos anos de 1940 e 1950. Seu objetivo era mudar o modo de acesso a essas instituições – e às posições de poder geralmente ocupadas pelos seus alunos –, de um sistema baseado na riqueza e classe social para um processo baseado principalmente na habilidade demonstrada por meio dos escores nos testes. Lemann argumenta que, embora este uso da testagem tenha aberto as portas do ensino superior aos filhos das classes média e baixa, ele também gerou uma nova elite meritocrática que se perpetua por gerações e em grande parte exclui os filhos das minorias raciais empobrecidas, que carecem das oportunidades educacionais precoces necessárias para o sucesso nos testes.

Testagem de pessoal e orientação vocacional

A utilização ótima dos talentos de cada pessoa é uma das principais metas da sociedade para a qual a testagem psicológica foi capaz de contribuir de um modo importante quase desde seu início. Decisões relativas à escolha vocacional precisam ser feitas em diferentes momentos da vida, geralmente durante a adolescência e o início da vida adulta, mas também cada vez mais na meia-idade, e as decisões quanto à seleção e colocação de pessoal nas empresas, indústrias e organizações militares precisam ser feitas de forma contínua. Alguns dos principais instrumentos que provaram ser particularmente úteis para este tipo de decisões são descritos nas seções a seguir.

Testes de aptidão e habilidades especiais

O sucesso do *Army Alpha* estimulou o interesse no desenvolvimento de testes com o objetivo de selecionar trabalhadores para diferentes ocupações. Ao mesmo tempo, profissionais da área da psicologia aplicada vinham desenvolvendo e empregando um conjunto básico de procedimentos que pudesse justificar o uso de testes na seleção ocupacional. Basicamente, os procedimentos envolviam:

a) identificar as *habilidades* necessárias para um determinado papel ocupacional por meio de uma *análise do trabalho*;
b) administrar testes elaborados para avaliar a *aptidão*;
c) correlacionar os resultados dos testes e mensurações do desempenho no trabalho.

Usando variações desse procedimento, a partir da década de 1920 os psicólogos foram capazes de desenvolver instrumentos para selecionar estagiários em campos tão diversos quanto o trabalho mecânico e a música. Teste de habilidades espaciais, motoras e organizacionais logo se seguiram. O campo da seleção de pessoal na indústria e nas Forças Armadas se desenvolveu em torno desses instrumentos, juntamente com o uso de amostras de tarefas, dados biográficos e testes gerais de inteligência, individuais e grupais. Muitos destes instrumentos também foram usados com êxito para identificar os talentos de jovens em busca de orientação vocacional.

Baterias de aptidões múltiplas

Na década de 1940, o uso de testes de habilidades separadas no aconselhamento vocacional seria substituído, em grande parte, por baterias de aptidões múltiplas, desenvolvidas por meio das técnicas de análise fatorial propostas por Spearman e aperfeiçoadas na Inglaterra e nos Estados Unidos ao longo dos anos de 1920 e 1930. Estas baterias são grupos de testes unidos por um formato e uma base de

pontuação comuns, que tipicamente definem um perfil dos pontos fortes e fracos do indivíduo, oferecendo escores separados em vários fatores como raciocínio verbal, numérico, espacial e lógico e habilidades mecânicas, em vez de um único escore global como os produzidos pelos testes de QI Binet e do exército. As baterias de aptidões múltiplas foram criadas depois que as análises fatoriais dos dados de testes de habilidade deixaram claro que a inteligência não é um conceito unitário, e que as habilidades humanas englobam uma ampla gama de componentes ou fatores separados e relativamente independentes.

Mensuração de interesses

Assim como os testes de aptidões e habilidades especiais surgiram na indústria e depois encontraram uso no aconselhamento vocacional, as mensurações de interesses foram criadas para fins de orientação vocacional e, mais tarde, foram empregadas na seleção de pessoal. Em 1914, Truman L. Kelley produziu um Teste de Interesse simples, possivelmente o primeiro inventário de interesses criado, com itens relativos a preferências em materiais de leitura e atividades de lazer, bem como alguns que envolviam conhecimento de palavras e informações gerais. No entanto, a revolução nesta área particular da testagem aconteceu em 1924, quando M. J. Ream elaborou uma chave empírica que diferenciava as respostas de vendedores bem-sucedidos e malsucedidos no Inventário de Interesses Carnegie, desenvolvido por Yoakum e seus alunos no Instituto de Tecnologia Carnegie em 1921 (DuBois, 1970). Este evento marcou o início de uma técnica conhecida como *chave empírica de critério*, que, após aperfeiçoamentos como procedimentos de validação cruzada e extensões para outras ocupações, seria usada no *Strong Vocational Interest Blank* (SVIB), publicado pela primeira vez em 1927, e em outros tipos de inventários. A versão atual do SVIB – denominada Inventário de Interesses Strong (SII, *Strong Interest Inventory*®) – é um dos inventários de interesses mais usados no mundo e foi seguida de um grande número de instrumentos deste tipo.

Testagem clínica

No início do século XX, o campo da psiquiatria tinha adotado formas mais sistemáticas de classificar e estudar a psicopatologia. Estes avanços forneceram ímpeto para o desenvolvimento de instrumentos que ajudassem a diagnosticar problemas psiquiátricos. Os principais exemplos deste tipo de ferramentas são discutidos aqui.

Inventários de personalidade

O primeiro dispositivo deste tipo foi a Lista de Dados Pessoais Woodsworth (P-D Sheet, *Woodsworth Personal Data Sheet*), um questionário desenvolvido durante a Primeira Guerra Mundial para fazer a triagem de recrutas que pudessem sofrer de

doenças mentais. Consistia em 116 afirmações a respeito de sentimentos, atitudes e comportamentos obviamente indicativos de psicopatologia, às quais o testando respondia simplesmente sim ou não. Embora demonstrasse algum potencial, a guerra terminou antes que este instrumento fosse colocado em uso operacional. Depois da guerra, houve um período de experimentação com outros itens menos óbvios e escalas delineadas para acessar neuroticismo, traços de personalidade – como introversão e extroversão – e valores. Também foram instituídas inovações na apresentação dos itens com o objetivo de reduzir a influência da desejabilidade social, como a técnica de escolha forçada introduzida no Estudo de Valores Allport-Vernon em 1931. No entanto, o inventário de personalidade mais bem sucedido daquela época, e que sobrevive até hoje, foi o Inventário Multifásico Minnesota da Personalidade (MMPI; Hathaway e McKinley, 1940). O MMPI combinava itens da *P-D Sheet* e outros inventários, mas usava a técnica da chave empírica de critério introduzida pelo SVIB. Esta técnica resultou em um instrumento menos transparente, que os testandos não podiam manipular tão facilmente porque muitos itens não faziam referência óbvia a tendências psicopatológicas.

Desde a década de 1940, os inventários de personalidade tiveram grande desenvolvimento. Muitos refinamentos foram introduzidos em sua construção, incluindo o uso de perspectivas teóricas – como o sistema de necessidades de Henry Murray (1938) – e métodos de consistência interna na seleção de itens. Além disso, a análise fatorial, que havia sido tão crucial para o estudo e a diferenciação de habilidades, também começou a ser usada no desenvolvimento de inventários de personalidade. Na década de 1930, J. P. Guilford foi o pioneiro no uso da análise fatorial para agrupar itens em escalas homogêneas, enquanto que nos anos de 1940 R. B. Cattell aplicou a técnica para tentar identificar os traços de personalidade mais essenciais e, por isso, merecedores de investigação e avaliação. Atualmente, a análise fatorial tem um papel integrante na maioria das facetas da teoria e na construção de testes.

Técnicas projetivas

Embora os inventários de personalidade tivessem algum sucesso, os profissionais de saúde mental que trabalhavam com populações psiquiátricas ainda sentiam necessidade de auxílio no diagnóstico e tratamento das doenças mentais. Na década de 1920, surgiu um novo gênero de ferramenta para a avaliação da personalidade e da psicopatologia. Estes instrumentos, conhecidos como *técnicas projetivas*, tinham suas raízes nos métodos de associação livre introduzidos por Galton e usados clinicamente por Kraepelin, Jung e Freud. Em 1921, um psiquiatra suíço chamado Hermann Rorschach publicou um teste que consistia em 10 manchas de tinta que deveriam ser apresentadas uma de cada vez ao examinando para que ele as interpretasse. A chave para o sucesso desta primeira técnica projetiva formal é que ela oferecia um método padronizado para obter e interpretar as reações aos cartões com as manchas de tinta, respostas que – de modo geral – refletem o modo singular do sujeito de perceber o mundo e se relacionar com ele. O Teste de Rorschach foi adotado por vários psicólogos americanos e propagado em várias universidades

e clínicas dos Estados Unidos depois de sua morte prematura em 1922. A técnica Rorschach, juntamente com outros instrumentos pictóricos, verbais e não-verbais, como o Teste de Apercepção Temática, testes de completamento de sentenças e desenhos da figura humana, vieram fornecer um repertório novo de ferramentas – mais sutis e incisivas do que os questionários – para a investigação de aspectos da personalidade, que os testandos podiam não ser capazes de revelar ou não estar dispostos a isso. Embora haja muita controvérsia a respeito de sua validade, basicamente porque costumam se valer de interpretações qualitativas tanto ou mais do que de escores numéricos, as técnicas projetivas ainda são uma parte significativa do repertório de muitos clínicos (Viglione e Rivera, 2003).

Testes neuropsicológicos

O papel da disfunção cerebral nos transtornos emocionais, cognitivos e comportamentais tem sido crescentemente reconhecido ao longo do último século. No entanto, o principal estímulo para o estudo científico e clínico das relações entre cérebro e comportamento, que são o objeto de estudo da *neuropsicologia*, veio das investigações de Kurt Goldstein sobre as dificuldades observadas em soldados que tinham sofrido lesões cerebrais durante a Primeira Guerra Mundial. Muitas vezes, estes soldados exibiam um padrão de déficits envolvendo problemas de pensamento abstrato, memória e planejamento e execução de tarefas relativamente simples, todos passaram a ser incluídos na rubrica da *organicidade*, palavra usada como sinônimo de dano cerebral. Ao longo de várias décadas, foram criados diversos instrumentos para detectar a organicidade e distingui-la de outros transtornos psiquiátricos. Muitos destes eram variações dos testes de desempenho não-verbais desenvolvidos para avaliar a capacidade intelectual geral de indivíduos que não podiam ser examinados em inglês ou que tinham problemas de fala ou audição. Estes testes envolviam materiais como tabuleiros de formas, quebra-cabeças e blocos, bem como tarefas de lápis e papel como labirintos e desenhos. Muito se aprendeu a respeito do cérebro e seu funcionamento nas últimas décadas, e grande parte das idéias iniciais em avaliação neuropsicológica teve que ser revisado a partir dessas novas informações. Os danos cerebrais não são mais vistos como uma condição "tudo ou nada" de organicidade com um conjunto comum de sintomas, mas sim como uma ampla gama de transtornos possíveis resultantes da interação de fatores genéticos e ambientais específicos em cada caso individual. Não obstante, o campo da avaliação neuropsicológica continua a crescer em número e tipos de instrumentos disponíveis, e contribui para a compreensão clínica e científica das muitas relações entre o funcionamento cerebral e a cognição, as emoções e o comportamento (Lezak, 1995).

USOS ATUAIS DOS TESTES PSICOLÓGICOS

Atualmente, as testagens de modo geral são mais sofisticadas metodologicamente e embasadas de forma mais consistente do que em qualquer época do passado. O

uso atual dos testes, que acontece em uma ampla variedade de situações, pode ser classificado em três categorias: (a) tomada de decisões, (b) pesquisas psicológicas e (c) autoconhecimento e desenvolvimento pessoal. Como indica esta lista, apresentada no quadro Consulta Rápida 1.5, os três tipos de uso diferem vastamente em seu impacto e em muitos outros aspectos, e o primeiro deles é certamente o mais visível ao público.

CONSULTA RÁPIDA 1.5

Usos atuais dos testes psicológicos

- A primeira e mais importante modalidade de uso dos testes ocorre no processo pragmático de *tomada de decisões a respeito de pessoas*, sejam indivíduos ou grupos.
- A segunda modalidade em termos da freqüência e longevidade está na *pesquisa científica sobre fenômenos psicológicos* e diferenças individuais.
- O uso mais recente e menos desenvolvido dos testes ocorre nos processos terapêuticos de *promoção ou autoconhecimento* e do *ajustamento psicológico*.

Tomada de decisões

O uso primário dos testes psicológicos ocorre como ferramenta para a tomada de decisões. Esta particular aplicação da testagem invariavelmente envolve julgamentos de valor por parte de uma ou mais pessoas que tomam as decisões e precisam determinar critérios para selecionar, alocar, classificar, diagnosticar ou conduzir outros processos com indivíduos, grupos, organizações ou programas. Naturalmente, este uso da testagem muitas vezes é carregado de controvérsia, pois costuma resultar em conseqüências desfavoráveis para uma ou mais partes. Em muitas situações nas quais pessoas envolvidas discordam das decisões finais, o uso dos testes em si é atacado, independentemente de ser ou não apropriado.

Quando são usados testes para a tomada de decisões significativas a respeito de indivíduos ou programas, a testagem deve ser meramente parte de uma estratégia bem-planejada e minuciosa, que leve em consideração o contexto particular no qual as decisões são tomadas, as limitações dos testes e outras fontes de dados além destes. Infelizmente, com muita freqüência – por pressa, descuido ou falta de informação – os testes são considerados os únicos responsáveis por falhas em processos de tomada de decisão que atribuem um peso excessivo aos seus resultados e negligenciam outras informações pertinentes. Um grande número de decisões rotineiras tomadas por instituições educacionais, governamentais ou empresariais, que geralmente envolvem a avaliação simultânea de várias pessoas, foram e ainda são elaboradas desta forma. Embora tenham conseqüências importantes – como contratações, admissão em universidades ou escolas profissionalizantes, formatura ou licenciamento para a prática profissional – para os indivíduos envolvidos, as decisões se baseiam quase exclusivamente nos escores de testes. Os profissionais

da área da testagem, bem como alguns órgãos governamentais, estão tentando mudar esta prática, que é um legado do modo como os testes se originaram. Um entre vários passos importantes nesta direção é a publicação de um guia de recursos para educadores e responsáveis pela formulação de políticas educacionais voltado para o uso de testes como parte da tomada de decisões críticas envolvendo estudantes (U.S. Department of Education, Office for Civil Rights, 2000).

Pesquisas psicológicas

Os testes muitas vezes são usados em pesquisas no campo da psicologia diferencial, evolutiva, educacional, social e vocacional, da psicopatologia, entre outros. Eles oferecem um método reconhecido para o estudo da natureza, do desenvolvimento e das inter-relações de traços cognitivos, afetivos e comportamentais. Na verdade, embora vários testes que tiveram origem no curso de investigações psicológicas tenham se tornado disponíveis comercialmente, um número muito maior de instrumentos permanecem arquivados em dissertações, periódicos e vários compêndios de mensuração experimental discutidos na seção *Fontes de Informações Sobre Testes* no final deste capítulo. Como raramente existem conseqüências práticas imediatas do uso de testes em pesquisas, sua aplicabilidade neste contexto é menos polêmica do que quando são usados na tomada de decisões a respeito de indivíduos, grupos, organizações ou programas.

Autoconhecimento e desenvolvimento pessoal

A maior parte dos psicólogos e conselheiros humanistas, muitas vezes com razão, considera que o campo da testagem dá uma ênfase exagerada à rotulação e à categorização dos indivíduos em termos de critérios numéricos rígidos. A partir dos anos de 1970, alguns destes profissionais, especialmente Constance Fischer (1985/1994), começaram a usar testes e outras ferramentas de avaliação de forma individualizada, consoante com os princípios humanistas e existencialistas-fenomenológicos. Esta prática, que considera a testagem uma forma de oferecer aos clientes informações que podem promover a autoconhecimento e o crescimento positivo, evoluiu para o *modelo terapêutico de avaliação* esposado por Finn e Tonsager (1997). Obviamente, a aplicação mais pertinente deste modelo acontece no aconselhamento e na psicoterapia, nos quais o cliente é o único usuário dos resultados dos testes.

AVALIAÇÃO PSICOLÓGICA *VERSUS* TESTAGEM PSICOLÓGICA

Por motivos em grande parte relacionados à comercialização dos testes, alguns autores e editoras começaram a usar a palavra *avaliação* nos títulos de seus testes. Por isso, aos olhos do público leigo os termos *avaliação* e *testagem* muitas vezes são sinônimos, o que é um fato lamentável. Muitos profissionais desta área acreditam

que a distinção entre os termos deve ser preservada e esclarecida para o público em geral, uma vez que essas pessoas são possíveis clientes de avaliações ou consumidores de testes.

O uso de testes para a tomada de decisões a respeito de uma pessoa, um grupo ou um programa sempre deve acontecer dentro do contexto de uma *avaliação psicológica*. Este processo pode ocorrer em serviços de saúde, no aconselhamento ou em procedimentos forenses, bem como no contexto educacional e profissional. A avaliação psicológica é um *processo* flexível e não-padronizado, que tem por objetivo chegar a uma determinação sustentada a respeito de uma ou mais questões psicológicas através da coleta, avaliação e análise de dados apropriados ao objetivo em questão (Maloney e Ward, 1976).

Não esqueça

- *Testes* e *avaliações* NÃO são sinônimos.
- Os testes são uma das ferramentas usadas no processo de avaliação.

Passos no processo de avaliação

O primeiro e mais importante passo na avaliação psicológica é identificar seus objetivos do modo mais claro e realista possível. Sem objetivos claramente definidos e acordados entre o avaliador e a pessoa que solicita a avaliação, o processo dificilmente será satisfatório. Na maioria dos casos, a avaliação termina com um relatório verbal ou escrito comunicando as conclusões às pessoas que solicitaram a avaliação, em formato útil e compreensível. Entre estes dois pontos, o profissional que conduz a avaliação, geralmente um psicólogo ou conselheiro, vai precisar empregar seus conhecimentos especializados em vários momentos. Esses passos envolvem a seleção apropriada dos instrumentos a serem usados na coleta de dados, a cuidadosa administração, pontuação e interpretação e – o mais importante – o uso criterioso dos dados coletados para fazer inferências a respeito da questão proposta. Este último passo vai além dos procedimentos psicométricos e requer o conhecimento da área à qual a questão se refere, como serviços de saúde, colocação educacional, psicopatologia, comportamento organizacional ou criminologia, entre outros. Exemplos de questões que se prestam à investigação através da avaliação psicológica incluem:

- *questões diagnósticas*, como diferenciar entre depressão e demência;
- *predições*, como estimar a probabilidade de comportamento suicida ou homicida;
- *julgamentos avaliativos*, como os envolvidos em decisões sobre a guarda de crianças ou na avaliação da eficácia de programas ou intervenções.

Nenhuma destas questões complexas pode ser resolvida somente por meio de escores de testes, pois uma mesma pontuação pode ter sentidos diferentes, depen-

dendo do examinando e do contexto no qual foi obtida. Além disso, nenhum escore de teste consegue captar todos os aspectos que precisam ser considerados ao se resolver essas questões.

Os testes psicológicos podem ser componentes-chave da avaliação psicológica, mas os dois processos diferem fundamentalmente em aspectos importantes. O quadro Consulta Rápida 1.6 lista diversas dimensões que diferenciam a testagem da avaliação psicológica. Embora exista pouca dúvida quanto à superioridade geral da avaliação em relação à testagem no que diz respeito à abrangência e utilidade, a maior complexidade do processo de avaliação torna seus resultados bem mais difíceis de operacionalizar do que os da testagem. Não obstante, mais recentemente começaram a ser reunidas evidências da eficácia da avaliação, pelo menos no campo dos serviços de saúde (Eisman et al., 2000; Kubiszyn et al., 2000; Meyer et al., 2001).

QUALIFICAÇÕES DOS USUÁRIOS DE TESTES

À medida que o número de testes continuou a crescer e seus usos se expandiram, não apenas nos Estados Unidos mas no mundo todo, a questão do seu mau uso despertou interesse crescente no público, no governo e em profissionais diversos. A psicologia, a profissão a partir da qual os testes surgiram e com a qual estão mais diretamente associados, assumiu a liderança na tentativa de combater seu mau uso. Os *Padrões de Testagem* promulgados pela APA e outras organizações profissionais (AERA, APA e NCME, 1999) são um importante veículo para este fim. A APA também aborda questões relacionadas à testagem e avaliação em seus princípios éticos e código de conduta (APA, 2002), assim como outras associações profissionais (p. ex., American Counseling Association, 1995; National Association of School Psychologists, 2000).

Embora as qualidades técnicas de vários testes estejam longe do ideal e possam contribuir para problemas em seu uso, de modo geral se admite que o motivo básico para o mau uso dos testes reside no conhecimento ou competência insuficientes por parte de muitos usuários. Os testes podem parecer relativamente simples e diretos para usuários em potencial que não estão cientes dos cuidados necessários em sua aplicação. Por causa disso, nas últimas décadas, associações profissionais dos Estados Unidos e outros países têm desenvolvido documentos que delineiam mais clara e especificamente do que antes as habilidades e a base de conhecimentos necessárias para um uso competente de testes (American Association for Counseling and Development, 1988; Eyde, Moreland, Robertson, Primoff e Most, 1988; International Test Commission, 2000; Joint Committee on Testing Practices, 1988).

Uma das exposições mais claras destes requisitos se encontra em um relatório preparado ao longo de cinco anos pela Força-Tarefa sobre Qualificações do Usuário de Testes da APA (APA, 2000). Este relatório delineia: (a) o conhecimento e as habilidades essenciais para o emprego de testes na tomada de decisões ou formulação de políticas que afetem a vida dos testandos e (b) os conhecimentos especializados que os usuários de testes nos contextos específicos de emprego, educação, aconselhamento profissional, serviços de saúde e tarefas forenses devem possuir.

CONSULTA RÁPIDA 1.6

Diferenças típicas entre testagem e avaliação psicológica

Aspecto	Testagem psicológica	Avaliação psicológica
Grau de complexidade	Mais simples; envolve um procedimento uniforme, freqüentemente unidimensional.	Mais complexa; cada avaliação envolve vários procedimentos (entrevistas, observações, testagens, etc.) e dimensões.
Duração	Mais breve, de alguns minutos a algumas horas.	Mais longa, de algumas horas a alguns dias ou mais.
Fontes de dados	Uma pessoa, o testando.	Muitas vezes são usadas fontes colaterais, como parentes ou professores, além do sujeito da avaliação.
Foco	Como uma pessoa ou grupo se compara com outros (*nomotético*).	A singularidade de um determinado indivíduo, grupo ou situação (*idiográfico*).
Qualificações necessárias	Conhecimento sobre testes e procedimentos de testagem.	Conhecimento de testagem e outros métodos de avaliação, bem como da área avaliada (p. ex., transtornos psiquiátricos, requisitos para uma função).
Base de procedimentos	É necessária objetividade; a quantificação é crucial.	É necessária subjetividade, na forma de julgamento clínico; a quantificação raramente é possível.
Custo	Barata, especialmente quando feita em grupos.	Muito cara, pois requer o uso intensivo de profissionais altamente qualificados.
Objetivo	Obter dados para uso na tomada de decisões.	Chegar a uma decisão a respeito da questão ou problema que originou o encaminhamento.
Grau de estruturação	Altamente estruturada.	Engloba aspectos estruturados e não-estruturados.
Avaliação dos resultados	Investigação relativamente simples da fidedignidade e validade baseada em resultados grupais.	Muito difícil devido à variabilidade de métodos, avaliadores, natureza das questões investigadas, etc.

Conhecimentos e habilidades genéricos em psicometria, estatística, seleção de testes, administração, pontuação, comunicação de resultados e salvaguardas são considerados relevantes para todos os usuários. Os conhecimentos adicionais e a experiência supervisionada necessários para o uso de testes nos vários contextos e com diversos grupos de testandos também são delineados no relatório, assim como os vários usos dos testes para fins de classificação, descrição, predição, planejamento de intervenções e rastreamento em cada um deles.

Outro aspecto da testagem que tem contribuído para o mau uso dos testes ao longo das décadas é a relativa facilidade com que os instrumentos podem ser obtidos por pessoas que não estão qualificadas a usá-los. Em certo grau, a disponibilidade dos testes é uma função da liberdade com a qual a informação flui em sociedades democráticas como a dos Estados Unidos, especialmente na era da internet. Outro motivo para este problema – já citado neste capítulo – é o fato de que muitos testes são produtos comerciais. Como resultado, algumas editoras estão dispostas a vendê-los para pessoas ou instituições sem observar as salvaguardas apropriadas para se certificarem de que eles possuem as credenciais corretas. Durante as décadas de 1950 e 1960, os *Padrões de Testagem* incluíam um sistema de três níveis para a classificação de testes em termos das qualificações necessárias para seu uso (APA, 1966, p.10-11). Este sistema, que encaixava os testes nos níveis A, B ou C dependendo da formação requerida para seu uso, era facilmente burlado por indivíduos em escolas, órgãos governamentais e empresas. Embora muitas editoras ainda usem este sistema, os *Padrões de Testagem* não mais o adotam. O quadro Consulta Rápida 1.7 delineia os elementos tipicamente incluídos no sistema triplo de classificação das qualificações de usuários de testes.

Em 1992, várias editoras de testes e prestadoras de serviços de avaliação criaram a Associação de Editoras de Testes (ATP, *Association of Test Publishers*). Esta organização sem fins lucrativos tem como objetivo manter um alto nível de profissionalismo e ética nas iniciativas de testagem. Uma de suas formas de monitorar a distribuição dos testes é através da exigência de alguma documentação que ateste um nível mínimo de formação daqueles que compram seus produtos.

Formulários de qualificação para a compra de testes agora são incluídos nos catálogos de todas as editoras respeitáveis. Por mais sinceros que sejam estes esforços para preservar a segurança dos materiais e impedir seu mau uso, sua eficácia é necessariamente limitada. Não apenas é impossível verificar nos formulários as qualificações que os compradores afirmam ter, como tampouco qualquer conjunto formal de qualificações – seja por formação ou licenciamento – pode garantir que um indivíduo seja efetivamente competente para usar um teste de modo adequado em uma determinada situação (ver Capítulo 7).

FONTES DE INFORMAÇÕES A RESPEITO DE TESTES

Na testagem psicológica, assim como em todas as outras atividades humanas, a internet criou um suprimento interminável de informações. Por isso, juntamente com as referências impressas, tradicionalmente encontradas nesta área, agora existe um grande número de recursos *on-line* e eletrônicos facilmente acessíveis.

CONSULTA RÁPIDA 1.7

Níveis de qualificação do usuário de testes

Todas as respeitáveis editoras de testes exigem que seus clientes preencham um formulário especificando as credenciais que os qualificam a usar os materiais que desejam comprar e certificando que eles serão usados de acordo com todas as diretrizes éticas e legais aplicáveis. Embora o número de níveis e as credenciais específicas exigidas em cada um deles variem entre as editoras, seus critérios de qualificação são tipicamente organizados em pelo menos três níveis, baseados em uma categorização dos testes e dos requisitos de formação delineada originalmente pela Associação Americana de Psicologia (APA, 1953, 1954).

	Nível inferior (A)	Nível intermediário (B)	Nível superior (C)
Tipo de instrumentos ao qual este nível se aplica	Uma gama limitada de instrumentos, como testes de realização educacional, que podem ser administrados, pontuados e interpretados sem treinamento especializado, seguindo-se as instruções dos manuais.	Ferramentas que exigem algum treinamento especializado na construção e uso de testes e na área na qual os instrumentos serão aplicados, como testes de aptidão e inventários de personalidade aplicáveis a populações normais	Instrumentos que requerem extensa familiaridade com princípios de testagem e avaliação, bem como com os campos psicológicos aos quais os instrumentos pertencem, como testes de inteligência individual e técnicas projetivas.
Credenciais ou requisitos necessários para a compra de materiais deste nível	Algumas editoras não exigem credenciais para a compra de testes deste nível. Outras podem exigir bacharelado na área específica ou solicitar que os pedidos de materiais sejam feitos através de uma órgão ou instituição.	Os compradores dos testes geralmente devem ter grau de mestre em psicologia (ou algum campo afim) ou experiência supervisionada em testagem e avaliação condizente com os requisitos para o uso dos instrumentos em questão.	Os compradores dos testes devem ter o tipo de formação avançada e experiência supervisionada que é adquirida no curso da obtenção do grau de doutorado, ou licenciamento profissional em um campo pertinente ao uso pretendido dos instrumentos, ou ambos.

Recursos na internet

Para a pessoa que busca informações a respeito de testes psicológicos, um bom ponto de partida é a seção de Testagem e Avaliação no *site* da APA (http://www.apa.org). Dentro dessa seção, entre outras coisas, existe um excelente artigo sobre "Perguntas mais freqüentes/Como encontrar informações a respeito de testes psicológicos" (APA, 2003), que oferece orientação sobre como localizar testes

publicados e inéditos bem como documentos importantes pertinentes à testagem psicológica. Os *testes publicados* estão disponíveis comercialmente por meio de editoras, embora às vezes possam estar esgotados como os livros. Os *testes inéditos* devem ser obtidos diretamente do investigador que os criou, a menos que apareçam nos periódicos científicos ou em diretórios especializados (discutido mais adiante).

Não esqueça

Uma das distinções mais básicas entre os testes diz respeito à existência de publicação ou não.

- *Testes publicados* estão disponíveis comercialmente através de editoras.
- *Testes inéditos* devem ser obtidos do investigador que os desenvolveu, em diretórios especiais de mensurações inéditas ou nos periódicos científicos.

Dois outros grandes pontos de entrada na Internet para quem busca informações sobre um teste específico são: (a) a página de Revisões de Testes *On-line* do Instituto Buros de Mensuração Mental (BI) (http://www.unl.edu/buros), que oferece informações gratuitas sobre quase 4 mil testes disponíveis comercialmente, bem como mais de 2 mil revisões que podem ser compradas para leitura *on-line* e (b) a base de dados da Relação de Testes do Serviço de Testagem Educacional (ETS) (http://www.ets.org/testcoll/index.html), que é a maior do seu tipo no mundo. Além disso, a página do Centro de Informações sobre Recursos Educacionais (ERIC) (http://eric.ed.gov) – mantida pelo Ministério da Educação dos Estados Unidos – contém uma grande quantidade de materiais relacionados à testagem psicológica.

Outra forma de obter informações *on-line* sobre testes publicados e inéditos é por meio de índices eletrônicos dos periódicos científicos em psicologia, educação ou administração. A base de dados PsycINFO da APA, disponível em muitas bibliotecas ou por assinatura, permite que se encontrem referências bibliográficas, resumos e até mesmo textos completos de artigos sobre um teste a partir de seu nome. Além dos títulos exatos, a PsycINFO e outras bases de dados podem ser pesquisadas por tema, palavra-chave e autor, o que as torna especialmente úteis quando só estão disponíveis informações parciais.

Não esqueça

O Apêndice A lista todos os testes e instrumentos de avaliação psicológica publicados disponíveis comercialmente que são mencionados ao longo deste livro, juntamente com os códigos que identificam suas editoras.

O Apêndice B fornece os endereços eletrônicos das editoras listadas no apêndice A. Informações mais detalhadas sobre editoras de testes, incluindo endereços reais e números de telefone, estão disponíveis na edição mais recente do *Tests in Print* (Murphy, Plake, Impara e Spies, 2002).

Depois que um teste é localizado através de qualquer um desses recursos, geralmente também se pode determinar se ele foi publicado e como pode ser obtido. Se o teste foi publicado, pode ser comprado da companhia que o publica por pessoas que satisfaçam as qualificações para usá-lo. As instruções para a compra constam dos catálogos das editoras, muitos deles agora estão disponíveis *on-line* bem como em formato impresso. O *site* da ATP (http://www.testpublishers.org) tem *links* para muitas editoras e prestadoras de serviços de avaliação. Os endereços eletrônicos de todas as organizações mencionadas nesta seção e outras fontes importantes de informações sobre testes podem ser encontrados no quadro Consulta Rápida 1.8.

CONSULTA RÁPIDA 1.8 Fontes de informação sobre testes psicológicos na internet

Organização (Sigla)	Endereço eletrônico
American Educational Research Association (ERA)	http://www.aera.net
American Psychological Association (APA)	http://www.apa.org
Association of Test Publishers (ATP)	http://www.testpublishers.org
Buros Institute of Mental Measurements (BI)	http://www.unl.edu/buros
Educational Resources Information Center (ERIC)	http://eric.ed.gov
Educational Testing Service (ETS)	http://www.ets.org/testcoll/index.html
International Test Commission (ITC)	http://www.intestcom.org
National Council on Measurement in Education (NCME)	http://www.ncme.org

Recursos impressos

Testes publicados

No que diz respeito aos testes publicados disponíveis comercialmente, as fontes mais importantes de informação estão ligadas ao Instituto Buros de Mensuração Mental (BI), sediado em Lincoln, Nebraska. Em particular, o BI (http://www.unl.edu/buros) produz duas séries de volumes que podem orientar os usuários de quase todos os testes publicados disponíveis nos Estados Unidos. Um deles é a série *Tests in Print (TIP)*, e o outro é a série *Mental Measurements Yearbook* (MMY). O *Tests in Print* é uma bibliografia abrangente de todos os testes disponíveis comercialmente no momento em que um determinado volume da série é publicado. Cada entrada apresenta o título do teste, sua sigla, autor, editora, data de publicação e outras informações básicas sobre o assunto, bem como referências cruzadas para as revisões do teste em todos os *MMYs* disponíveis naquele momento. A série *TIP* contém um índice de classificação de testes extremamente útil, bem como índices de esco-

res, editoras, siglas e nomes dos autores e revisores. A série *MMY*, por sua vez, remonta a 1938, quando o falecido Oscar Buros publicou o primeiro volume para auxiliar os usuários de testes com revisões avaliativas escritas por profissionais qualificados e independentes. Embora os *MMYs* ainda sejam publicados em forma de livro, suas entradas e revisões também estão disponíveis *on-line* e em outros meios eletrônicos. O Instituto Buros também publica muitos outros materiais relacionados a testes.

A PRO-ED (http://www.proedinc.com) publica *Tests*, uma série de volumes enciclopédicos que traz descrições sucintas de instrumentos em psicologia, educação e administração. A série *Test Critiques*, que remonta a 1984, complementa a *Tests*. Cada volume desta série contém revisões de testes e índices cumulativos para todos os volumes anteriores.

Testes inéditos

O objetivo dos cientistas comportamentais que usam testes psicológicos é investigar constructos psicológicos e diferenças individuais e grupais. Muitos testes existentes são usados exclusivamente para pesquisas científicas e não estão disponíveis comercialmente. Esses testes são referidos como mensurações *inéditas*, porque não podem ser comprados; as condições para o seu uso são tipicamente estabelecidas pelos autores de cada instrumento e quase sempre demandam uma carta solicitando permissão para usá-los. Informações sobre testes inéditos – e muitas vezes os próprios instrumentos – estão disponíveis nos periódicos científicos em psicologia (p. ex., através da PsycINFO *on-line*) e em vários diretórios (p. ex., Goldman, Mitchell e Egelson, 1997; Robinson, Shaver e Wrightsman, 1991). O artigo mencionado anteriormente "Perguntas mais freqüentes/Como encontrar informações sobre testes psicológicos" (APA, 2003) lista diversos recursos impressos e eletrônicos para informações sobre testes inéditos.

Teste a si mesmo

1. Qual dos seguintes não é um elemento essencial da testagem psicológica?
   (a) procedimentos sistemáticos
   (b) uso de padrões derivados empiricamente
   (c) regras preestabelecidas para a pontuação
   (d) amostragem de comportamentos de domínios afetivos

2. A fonte mais importante de critérios para a avaliação de testes, práticas de testagem e efeitos do uso de testes pode ser encontrada em:
   (a) *princípios éticos dos psicólogos e código de conduta.*
   (b) *padrões de testagem educacional e psicológica*
   (c) *manual diagnóstico e estatístico dos transtornos mentais*
   (d) *relatório da força-tarefa sobre qualificações do usuário de testes*

3. Os primeiros antecedentes da moderna testagem para seleção de pessoal remontam a
   (a) China a.C.
   (b) Grécia antiga
   (c) Império inca
   (d) Europa medieval

4. A avaliação dos testes psicológicos é menos problemática
   (a) antes de eles serem colocados em uso
   (b) depois que eles foram colocados em uso

5. Comparado às outras áreas listadas, o desenvolvimento de critérios ou bases para a tomada de decisões tem sido substancialmente lento no contexto da
   (a) avaliação educacional
   (b) avaliação ocupacional
   (c) avaliação clínica

6. O crédito pela criação do primeiro teste psicológico bem sucedido na era moderna geralmente é atribuído a
   (a) Francis Galton
   (b) Alfred Binet
   (c) James McKeen Cattell
   (d) Wilhelm Wundt

7. O QI-razão ou quociente de inteligência foi derivado
   (a) somando-se a idade mental (IM) e a idade cronológica (IC) do testando
   (b) subtraindo-se a IC da IM e multiplicando-se o resultado por 100
   (c) dividindo-se a IC pela IM e multiplicando-se o resultado por 100
   (d) dividindo-se a IM pela IC e multiplicando-se o resultado por 100

8. O objetivo básico para o qual os testes psicológicos são usados atualmente é
   (a) pesquisa psicológica
   (b) pesquisa educacional
   (c) tomada de decisões
   (d) autoconhecimento e desenvolvimento pessoal

9. Comparada à testagem psicológica, a avaliação psicológica geralmente é
   (a) mais simples
   (b) mais estruturada
   (c) mais cara
   (d) mais objetiva

10. Qual das seguintes seria a melhor fonte de informação sobre um teste que não está disponível comercialmente?
   (a) *Mental Measurements Yearbook*
   (b) *Test Critiques*
   (c) *Tests in Print*
   (d) PsycINFO

Respostas: 1.d; 2.b; 3.a; 4.a; 5.c; 6.b; 7.d; 8.c; 9.c; 10.d.

[capítulo **2**]

ESTATÍSTICA BÁSICA PARA TESTAGEM

De modo geral o progresso da ciência é acompanhado da invenção de instrumentos de mensuração e avanços em seus procedimentos e técnicas. A ciência da astronomia, por exemplo, decolou realmente nos séculos XVII e XVIII após a invenção de um telescópio adequado à observação do cosmos e do desenvolvimento da geometria analítica por Descartes, que levou ao cálculo mais preciso das distâncias entre os corpos celestes, entre outras coisas. Igualmente, os enormes avanços atuais no campo da neurociência devem muito ao desenvolvimento de técnicas como a tomografia por emissão de posítron (PET) e a ressonância magnética funcional (RMf), que permitem aos cientistas visualizar e medir pequenas alterações e eventos bioquímicos no cérebro.

Como vimos no Capítulo 1, o moderno campo da testagem psicológica também teve seu início com a invenção de ferramentas bem-sucedidas. As escalas de inteligência Binet-Simon possibilitaram a mensuração de processos cognitivos importantes – como a compreensão, o julgamento e a memória – por meio de amostras de comportamento calibradas conforme a idade. A invenção dos itens objetivos de múltipla escolha por Arthur Otis levou aos primeiros testes grupais de inteligência geral. Técnicas estatísticas desenvolvidas aproximadamente na mesma época que os primeiros testes permitiram a análise dos dados coletados por meio destes.

A MENSURAÇÃO

O conceito de mensuração está no centro da testagem psicológica como atividade científica voltada para o estudo do comportamento humano. A mensuração envolve o uso de certos dispositivos ou regras para atribuir números a objetos ou eventos (Stevens, 1946). Se aplicarmos este processo sistematicamente, os fenômenos medidos estarão mais facilmente sujeitos à confirmação e análise e, portanto, tornar-se-ão mais objetivos. Em outras palavras, ao analisarmos, categorizarmos e quantificarmos sistematicamente os fenômenos observáveis, nós os trazemos para a arena científica.

É central para a definição dos testes psicológicos o fato de que os fenômenos consistem em amostras cuidadosamente escolhidas de comportamentos às quais é aplicado um sistema numérico ou categórico segundo alguns padrões preestabelecidos. A testagem psicológica é amplamente co-extensiva ao campo da *psicometria*, ou mensuração psicológica, e é uma das ferramentas primárias para a ciência e a prática da psicologia.

O uso de números na testagem requer que nos aprofundemos em estatística. Para muitos estudantes de psicologia, o uso de estatística e dados quantitativos em geral representa um problema que pode parecer insuperável: lidar com números tende a causar alguma ansiedade. Esta ansiedade está ligada às dificuldades muitas vezes encontradas nas aulas de matemática e estatística por motivos que podem estar relacionados tanto a fatores emocionais e comportamentais quanto aos temas em si ou ao modo como eles têm sido tradicionalmente ensinados. Este capítulo apresenta os conceitos estatísticos necessários para a compreensão dos princípios básicos da testagem psicológica. Os leitores que dominam a estatística básica talvez possam pular todo o capítulo ou a maior parte dele, mas quanto ao resto, qualquer leitor motivado deste livro pode obter uma compreensão instrumental dos conceitos que serão apresentados. É importante, no entanto, perceber que estes conceitos seguem uma progressão lógica: para passar a um novo tópico, é essencial dominar os precedentes. Um auxílio adicional na compreensão dos métodos estatísticos básicos está disponível em muitas obras excelentes, como aquelas listadas no quadro Consulta Rápida 2.1.

CONSULTA RÁPIDA 2.1

Conselhos sobre estatística

Premissas básicas

1. Para compreender os testes psicológicos, é preciso lidar com números e estatística.
2. Compreender a estatística é possível para qualquer leitor deste livro.
3. A melhor maneira de aumentar a compreensão dos conceitos estatísticos é através de sua aplicação.

Fontes de auxílio recomendadas

Livros

- Howell, D.C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.
- Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Fort Worth, TX: Harcourt Brace.
- Urdan, T. C. (2001). *Statistics in plain english*. Mahwah, NJ: Erlbaum.
- Vogt, W. P. (1998). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (2nd ed). Thousand Oaks, CA: Sage.

Vídeo

- Blatt, J. (Produtor/Autor/Diretor). (1989). *Against all odds: inside statistics* [fita VHS]. (À venda em The Annenberg/CPB Project, 901 E St., NW, Washington, DC 2004-2006)

VARIÁVEIS E CONSTANTES

Uma das distinções mais básicas que podemos fazer em qualquer ciência é entre variáveis e constantes. Como os próprios termos indicam, uma variável é qualquer coisa que varia, enquanto que uma constante é qualquer coisa que não varia. Nosso mundo tem muitas variáveis e poucas constantes. Um exemplo de constante é o p (pi), a razão entre a circunferência de um círculo e seu diâmetro, um número que geralmente é arredondado para 3.1416. As variáveis, por outro lado, estão em toda parte e podem ser classificadas de várias formas. Por exemplo, algumas variáveis são visíveis (p. ex., sexo, cor dos olhos) e outras invisíveis (p. ex., personalidade, inteligência); algumas são definidas de tal modo que dizem respeito a conjuntos muito pequenos, e outras a conjuntos muito grandes (p. ex., o número de filhos de uma família ou a renda média dos indivíduos de um país); algumas são contínuas, e outras, discretas.

Esta última distinção é importante para os nossos fins e merece uma explicação. Tecnicamente, variáveis *discretas* são aquelas com uma gama finita de valores – ou potencialmente infinita, porém contável. As variáveis *dicotômicas*, por exemplo, são variáveis discretas que podem assumir apenas dois valores, como o sexo ou o resultado de um lance de cara ou coroa. As variáveis *politômicas* são variáveis discretas que podem assumir mais de dois valores, como estado civil, raça, etc. Outras variáveis discretas podem assumir uma gama mais ampla de valores, mas ainda assim podem ser contadas como unidades separadas; exemplos destas são o tamanho de uma família, a contagem de tráfego veicular e resultados de *beisebol*. Embora na prática seja possível cometer erros na contagem, em princípio as variáveis discretas podem ser calculadas precisa e acertadamente.

Variáveis *contínuas* como tempo, distância e temperatura, por outro lado, têm variações infinitas e não podem realmente ser contadas. São medidas com escalas que teoricamente podem ser subdivididas até o infinito e não têm interrupções entre seus pontos, como as escalas de relógios analógicos, réguas e termômetros de vidro. Como nossos instrumentos de mensuração (mesmo os relógios atômicos!) não podem ser calibrados com precisão suficiente para medir variáveis contínuas com absoluta exatidão, as mensurações que fazemos delas são aproximações mais ou menos precisas.

Antes de começarmos a lidar com números, mais uma advertência é aconselhável. Na testagem psicológica, quase sempre estamos interessados em variáveis contínuas (p. ex., graus de integridade, extroversão ou ansiedade), mas nós as mensuramos com ferramentas, como testes ou inventários, que não são tão precisas quanto as das ciências físicas e biológicas. Mesmo nestas ciências, a mensuração discreta de variáveis contínuas apresenta algumas limitações quanto à precisão. Por isso, fica óbvio que nas ciências comportamentais devemos estar particularmente atentos para potenciais fontes de erros e procurar estimativas pertinentes de erro sempre que nos defrontarmos com os resultados de qualquer processo de mensuração. Por exemplo, se números extraídos de amostras de eleitores em potencial forem usados para estimar o resultado de uma eleição, as margens de erro estimadas devem ser divulgadas juntamente com os resultados da pesquisa.

Em resumo, quando examinamos os resultados de qualquer processo de mensuração, precisamos ter muito claro o fato de que eles são *inexatos*. Em relação à testagem psicológica em particular, sempre que os escores de um teste são relatados, o fato de eles serem estimativas deve ser explicitado. Além disso, os limites dentro dos quais os escores podem variar, bem como os níveis de confiança para estes limites, precisam ser divulgados, juntamente com informações interpretativas (ver Capítulo 4).

Não esqueça

- Embora os números possam parecer precisos, todas as mensurações estão sujeitas a erro.
- Quando medimos variáveis discretas, os erros ocorrem apenas na contagem incorreta. A boa prática requer a prevenção, detecção e correção da contagem inexata.
- Quando medimos variáveis contínuas, por outro lado, os erros de mensuração são inevitáveis, como conseqüência das limitações das ferramentas de mensuração.
- Como ferramentas de mensuração, os testes psicológicos estão sujeitos a muitas limitações. Por isso, as margens de erro sempre devem ser estimadas e comunicadas juntamente com os resultados do teste.

O SIGNIFICADO DOS NÚMEROS

Como os números podem ser usados de várias formas, S. S. Stevens (1946) criou um sistema para classificar diferentes níveis de mensuração a partir das relações entre os números e os objetos ou eventos aos quais são aplicados. Estes níveis de mensuração ou escalas – delineadas na Tabela 2.1 – especificam algumas das principais diferenças no modo como os números podem ser usados, bem como os tipos de operações estatísticas que são logicamente viáveis, dependendo de como os números são usados.

Escalas nominais

No nível mais simples de sua classificação, Stevens localizou o que denominou escalas *nominais*. A palavra *nominal* deriva do radical latino *nomem*, que significa *nome*. Como o termo sugere, nestas escalas os números são usados somente como rótulos para identificar um indivíduo ou classe. O uso nominal dos números para indicar indivíduos é exemplificado pelos números da previdência social (*SS#s*) que identificam a maioria das pessoas que vivem nos Estados Unidos. Estes números são úteis porque cada um deles é atribuído a apenas uma pessoa e, portanto, pode servir para identificá-la mais especificamente do que seu nome e sobrenome, que podem ser comuns a muitas pessoas. Os números também podem ser usados para rotular *dados categóricos**, que são dados relacionados a variáveis como gênero, afiliação política, cor, etc. – isto é, dados que derivam da designação de pessoas, objetos ou eventos a categorias ou classes em particular. Ao passar dados demográ-

*N. de R.T. São dados que representam as variáveis numéricas.

TABELA 2.1 Níveis de mensuração

Tipo de escala	Características definidoras	Propriedades dos números	Exemplos
Nominal	São usados números ao invés de palavras.	Identidade ou igualdade	Números de inscrição na previdência social; números das camisetas de jogadores de futebol, códigos numéricos para variáveis não- quantitativas como sexo ou diagnósticos psiquiátricos.
Ordinal	São usados números para ordenar uma série hierárquica.	Identidade + ordem de classificação	*Ranking* de atletas ou times, escores de percentil.
Intervalar	Intervalos iguais entre as unidades mas sem zero verdadeiro.	Identidade + ordem de classificação + igualdade de unidades	Escalas de temperatura Farenheit e Celsius, calendário.
Racional	O zero significa "nenhuma quantidade" do que é medido; todas as operações aritméticas são possíveis e significativas.	Identidade + ordem de classificação + igualdade de unidades + aditividade	Medidas de comprimento, períodos de tempo.

ficos para o computador para análise, por exemplo, os investigadores tipicamente criam uma escala nominal que usa números para indicar os níveis de uma variável categórica (ou seja, discreta). Por exemplo, o número 1 (um) pode ser atribuído a todas as mulheres, e o 2 (dois) a todos os homens. O único requisito para este uso dos números é que todos os membros de um conjunto designado por um determinado número devem ser iguais com respeito à categoria a ele atribuída. Naturalmente, embora os números usados nas escalas nominais possam ser somados, subtraídos, multiplicados e divididos, os resultados destas operações não são significativos. Quando usamos números para identificar categorias, como aprovado-reprovado, ou diagnósticos psiquiátricos, sua única propriedade é a *identidade*, o que significa que todos os membros de uma categoria devem receber o mesmo número e que duas categorias não podem compartilhar o mesmo número. A única operação aritmética permissível é a contagem das freqüências dentro de cada categoria. Obviamente, também é possível tratar estas freqüências calculando proporções e diversificando as análises baseadas nelas.

Escalas ordinais

Os números usados nas escalas *ordinais* comunicam uma pequena mas significativa informação a mais do que os números das escalas nominais. Nestas escalas,

além da identidade, existe a propriedade da *ordem de classificação*, o que significa que os elementos de um conjunto podem ser organizados em uma série – do mais baixo ao mais alto ou vice-versa – a partir de uma única variável, como ordem de nascimento ou nível de desempenho acadêmico dentro de uma determinada turma. Embora os números da ordem de classificação transmitam um sentido preciso em termos de posição, eles não informam a respeito da distância entre as posições. Portanto, os alunos de uma turma podem ser classificados em termos de seu desempenho, mas esta ordenação não vai refletir a quantidade de diferença entre eles, que pode ser grande ou pequena. Da mesma forma, em qualquer organização hierárquica como, digamos, a marinha dos Estados Unidos, os postos hierárquicos (p. ex., guarda-marinha, tenente, comandante, capitão, almirante) denotam posições diferentes, da mais baixa à mais alta, mas as diferenças entre elas em termos de realização ou prestígio não são as mesmas. Se a estas categorias fossem atribuídos números, tais como 1, 3, 7, 14 e 35, a ordem de precedência seria mantida, mas nenhum sentido adicional seria acrescentado.

Na testagem psicológica, o uso de números ordinais para comunicar resultados de testes é generalizado. Escores ordenados por classificação são relatados como *escores de postos de percentil* (PP) – que não devem ser confundidas com os escores percentuais amplamente usados em notas escolares. Os escores de percentil são simplesmente números ordinais dispostos em uma escala de 100 de tal modo que suas posições indicam a percentagem de indivíduos de um grupo que se enquadram em um determinado nível de desempenho ou abaixo dele. Por exemplo, o escore de posto de percentil de 70 indica um nível de desempenho igual ou superior ao de 70% das pessoas do grupo em questão. Os escores de posto de percentil, muitas vezes denominados simplesmente *percentis*, são o principal veículo pelo qual os usuários de testes transmitem as informações normativas derivadas dos testes e, por isso, serão discutidos mais detalhadamente no capítulo seguinte.

Os dados numéricos das escalas ordinais podem ser tratados estatisticamente da mesma forma que os dados nominais. Além disso, existem algumas técnicas estatísticas, como o coeficiente de correlação *rho* de Spearman (r_s) para diferenças de posto, que são apropriadas especificamente para uso com dados ordinais.

Escalas intervalares

Nas escalas intervalares, também conhecidas como *escalas de unidades iguais*, os números adquirem outra importante propriedade. Nelas, a diferença entre quaisquer dois números consecutivos reflete uma diferença empírica ou demonstrável igual entre os objetos ou eventos que os números representam. Um exemplo disso é o uso dos dias para marcar a passagem do tempo no calendário. Um dia consiste de 24 horas, cada uma delas com 60 minutos, e cada um destes com 60 segundos. Se duas datas estão separadas por 12 dias, elas estão exatamente três vezes mais distantes do que duas datas que têm apenas 4 dias de diferença. Observe, no entanto, que o tempo do calendário em meses não é uma escala de unidades iguais, porque alguns meses são mais longos do que outros. Além disso, o calendário tam-

bém tipifica uma característica das escalas intervalares que limita o sentido dos números usados nelas, qual seja, a inexistência de um ponto zero verdadeiro. No caso do calendário, não existe um ponto de partida para o início do tempo aceito por todos. Diferentes culturas determinam pontos de partida arbitrários, como o ano em que se acredita que Cristo nasceu, para marcar a passagem dos anos. Por exemplo, a tão antecipada chegada do novo milênio ao final do ano 2000 da era cristã ou comum ocorreu no ano de 5771 do calendário judaico e no ano de 4699 do calendário chinês, ambos começam muitos anos antes do início da era comum.

Nas escalas intervalares, as distâncias entre os números são significativas. Por isso, podemos aplicar a maioria das operações aritméticas a estes números e obter resultados que fazem sentido. No entanto, devido à arbitrariedade dos pontos-zero, os números de uma escala intervalar não podem ser interpretados em termos de razões.

Escalas de razão

Dentro das escalas *de razão*, os números adquirem a propriedade da aditividade, o que significa que eles podem ser somados – bem como subtraídos, multiplicados e divididos – e o resultado pode ser expresso como uma razão, com resultados significativos. As escalas de razão têm um ponto-zero verdadeiro ou absoluto que representa "nenhuma quantidade" do que está sendo medido. Nas ciências físicas, o uso desse tipo de escala de mensuração é comum: tempo, distâncias, pesos e volumes podem ser expressos como razões de uma forma significativa e logicamente consistente. Por exemplo, um objeto que pesa 16 libras é duas vezes mais pesado que um objeto que pesa 8 libras (16/8 = 2), assim como um objeto de 80 libras é duas vezes mais pesado que um objeto de 40 libras (80/40 = 2). Além disso, o ponto-zero na escala do peso indica a ausência absoluta de peso. Na psicologia, as escalas de razão são usadas primariamente quando tomamos medidas em termos de contagens de freqüência ou intervalo de tempo e ambos permitem a existência de zeros verdadeiros.

Dados categóricos ou discretos podem ser medidos – ou explicados – somente com escalas nominais, ou com escalas ordinais se os dados se encaixam em uma seqüência de algum tipo. Dados contínuos ou métricos podem ser medidos com escalas intervalares, ou escalas de razão se houver um ponto-zero verdadeiro. Além disso, dados contínuos podem ser convertidos em classes ou categorias e manipulados com escalas nominais ou ordinais. Por exemplo, podemos separar as pessoas em apenas três categorias – altas, médias e baixas –, estabelecendo dois pontos de corte arbitrários na variável contínua da altura.

Quando passamos das escalas nominais para as escalas de razão, vamos de números que transmitem menos informações a números que transmitem mais informações. Como conseqüência disto, passar de um nível de mensuração para outro requer que nos certifiquemos de que as informações que os números contêm sejam preservadas ao longo de todas as transformações ou manipulações que lhes forem aplicadas.

Não esqueça

- Na mensuração, deve haver um elo demonstrável entre os números aplicados a objetos, eventos ou pessoas e a realidade que estes números representam.
- Quando as regras usadas para criar este elo não são compreendidas, os resultados do processo de mensuração são facilmente mal interpretados.
- Quando passamos de um nível de mensuração para outro, devemos nos certificar de que as informações que os números contêm sejam preservadas nas transformações que aplicarmos.
- Os escores são números com sentidos específicos. Se as limitações destes sentidos não forem compreendidas, inferências equivocadas poderão ser feitas a partir dos escores.

Por que o sentido dos números é relevante para a testagem psicológica?

Embora não seja adotado universalmente, o sistema de Stevens para a classificação de escalas de mensuração ajuda a manter a *relatividade* do sentido dos números na perspectiva correta. Os resultados da maioria dos testes psicológicos são expressos em escores, que são números com sentidos específicos. Se as limitações destes sentidos não forem compreendidas, poderão ser feitas inferências equivocadas a partir dos escores. Infelizmente, isso é muito freqüente, como pode ser visto nos seguintes exemplos.

Exemplo 1: Limitações específicas das escalas ordinais. Como foi mencionado anteriormente, muitos escores são relatados na forma de postos de percentil, que são números de nível ordinal que não implicam igualdade de unidades. Se dois escores são separados por cinco unidades de postos de percentil – por exemplo, os percentis 45° e 50° – a diferença entre elas e o que esta diferença representa em termos do que está sendo mensurado não podem ser equacionadas com a diferença que separa quaisquer outros escores com diferença de cinco unidades de percentil – por exemplo, os percentis 90° e 95°. Em uma distribuição de escores que se aproxima da curva normal, discutida mais adiante neste capítulo e ilustrada no quadro Consulta Rápida 2.2, a maioria dos escores de teste se agrupam em torno do centro da distribuição. Isso significa que nestas distribuições as diferenças entre os escores de postos são sempre maiores nos extremos ou caudas da distribuição do que no meio.

Exemplo 2: O problema do QI-razão. Os quocientes de inteligência originais planejados para uso com a Escala de Inteligência Stanford-Binet (S-B) eram *QIs-razão*. Isto quer dizer que eles eram verdadeiros *quocientes*, derivados da divisão do escore de idade mental (IM) que a criança obtinha no teste S-B por sua idade cronológica (IC) e pela posterior multiplicação do resultado por 100 para eliminar os decimais. A idéia era de que as crianças médias teriam idades mentais e cronológicas semelhantes e QI de aproximadamente 100. Crianças com funcionamento abaixo

da média teriam idade mental mais baixa do que a cronológica e QI abaixo de 100, enquanto que as que funcionavam acima da média teriam idade mental mais alta do que a cronológica e QI acima de 100. Esta noção funcionava bastante bem para crianças até meados da idade escolar, um período durante o qual tende a haver um ritmo bastante uniforme de crescimento intelectual de um ano para o outro. No entanto, a razão IM/IC simplesmente não funcionava para adolescentes e adultos, porque seu desenvolvimento intelectual é muito menos uniforme – e as mudanças muitas vezes são imperceptíveis – de um ano para o outro. O fato de a idade cronológica máxima usada no cálculo do QI-razão da S-B original ser de 16 anos, independentemente da idade real da pessoa testada, criava problemas adicionais de interpretação. Além disso, as escalas de idade mental e cronológica não estão no mesmo nível de mensuração. A idade mental, como avaliada pelos primeiros testes de inteligência, era basicamente uma medida ordinal de nível, enquanto que a idade cronológica pode ser medida com uma escala de razão. Por esses motivos, dividir um número pelo outro para obter um quociente simplesmente não conduzia a resultados logicamente consistentes e significativos. O quadro Consulta Rápida 2.2 mostra exemplos numéricos que enfatizam alguns dos problemas que fizeram com que os QIs-razão fossem abandonados.

CONSULTA RÁPIDA 2.2

Exemplos de cálculos de QI-razão e problemas correspondentes

Sujeito	Idade mental (IM)	Idade cronológica (IC)	Diferença (IM-IC)	QI-razão[a]
Ally	6 anos	5 anos	1 ano	6/5 x 100 = 120
Ben	12 anos	10 anos	2 anos	12/10 x 100 = 120
Carol	18 anos	15 anos	3 anos	18/15 x 100 = 120

[a] No cálculo efetivo do QI-razão, a idade mental e a cronológica eram expressas em meses em vez de anos.

Problema 1: O escore de idade mental necessário para se obter um determinado QI continua a aumentar para cada idade cronológica sucessiva, de modo que os QIs-razão em diferentes idades cronológicas não são equivalentes.

Problema 2: Enquanto a idade cronológica aumenta uniformemente, o mesmo não acontece com a idade mental. Uma vez que a mais alta idade mental que pode ser alcançada em um determinado teste de inteligência não pode ser ilimitada, mesmo quando se estabelece um limite para a idade cronológica máxima usada para calcular o QI – como foi feito na escala S-B por muito tempo –, o QI que a maioria dos adultos pode obter é limitado artificialmente, se comparado ao de crianças e adolescentes.

Solução: Devido a estes e outros problemas do QI-razão, bem como ao conceito de idade mental, seu uso foi abandonado. O termo QI agora é usado para um escore que não é um QI-razão e nem mesmo um quociente. Este escore, conhecido como *QI de desvio*, foi introduzido por David Wechsler e é explicado no Capítulo 3.

O que podemos concluir sobre o sentido dos números nas mensurações psicológicas?

Na psicologia, é essencial ter em mente que a maioria das nossas escalas de mensuração são de natureza ordinal. A igualdade das unidades é aproximada pelas escalas usadas em muitos tipos de escores de teste, mas esta igualdade nunca é tão permanente ou completa quanto nas ciências físicas, porque as próprias unidades são relativas ao desempenho das amostras das quais são derivadas. O uso de escalas de razão na psicologia se limita a mensurações de freqüência, tempo de reação ou variáveis que podem ser expressas significativamente em unidades físicas. Por exemplo, se estivéssemos usando a produção por hora de uma linha de montagem como medida de velocidade do desempenho em uma função específica, poderíamos dizer que o Trabalhador A, que produz 15 unidades por hora, é três vezes mais rápido do que o Trabalhador B, que produz apenas 5 unidades por hora. Observe, no entanto, que não podemos dizer que o Trabalhador A é três vezes melhor do que o Trabalhador B, porque a velocidade provavelmente não é o único índice de desempenho no trabalho, mesmo em uma linha de montagem. O nível geral de desempenho é uma variável mais complexa, que provavelmente pode ser avaliada apenas com uma escala qualitativa ordinal.

TIPOS DE ESTATÍSTICA

Uma vez que o uso de números para representar objetos e eventos é tão generalizado na testagem psicológica, o trabalho nesta área envolve uma aplicação substancial da estatística, um ramo da matemática dedicado a organizar, representar, resumir, analisar e manipular de outras formas os dados numéricos. Os números e gráficos usados para descrever, condensar ou representar dados pertencem ao domínio da *estatística descritiva*. Por outro lado, quando são usados dados para estimar valores populacionais baseados em valores de amostras ou para testar hipóteses, é aplicada a *estatística inferencial* – um conjunto mais amplo de procedimentos baseados na teoria das probabilidades. Felizmente, embora tanto a estatística descritiva quanto a inferencial sejam usadas extensamente no desenvolvimento de testes, a maioria dos aspectos quantitativos da interpretação dos escores de teste requer apenas uma boa compreensão da estatística descritiva e um número relativamente menor de técnicas do tipo inferencial. Além disso, muito embora uma formação em matemática de alto nível seja desejável para a compreensão completa da estatística envolvida na testagem, é possível compreendê-la em nível básico com uma boa dose de lógica e um conhecimento relativamente limitado de matemática.

A palavra *estatística* também é usada para se referir a medidas derivadas de dados de amostras – diferentes das derivadas de populações, que são chamadas de *parâmetros*. Médias, desvios padrões, coeficientes de correlação e outros números calculados a partir de amostras de dados são estatísticas derivadas para se estimar o que realmente interessa, ou seja, os respectivos parâmetros populacionais. Parâmetros são números matematicamente exatos (ou constantes, como o p) que geral-

mente não são obteníveis a menos que uma população seja tão fixa e circunscrita que todos os seus membros possam ser incluídos, como todos os membros de uma turma universitária em um determinado semestre. Na verdade, um dos principais objetivos da estatística inferencial é estimar parâmetros populacionais a partir de dados de amostras e da teoria da probabilidade.

— Não esqueça

Os dois sentidos de Estatística

1. O estudo e a aplicação de métodos para organizar, representar, resumir, analisar e tratar de outras formas dados numéricos.
2. Números (p. ex., médias, coeficientes de correlação) que descrevem as características de variáveis ou conjuntos de dados derivados de amostras, em oposição aos derivados de populações, que são denominados *parâmetros*.

Estatística descritiva

Dados brutos não são muito úteis. Geralmente consistem em um grupo de números que não transmitem qualquer sentido, mesmo depois de mais de um exame aprofundado, como os 60 números listados na Tabela 2.2. Estes números são os escores de 60 estudantes universitários no primeiro teste (com 50 itens de múltipla escolha) aplicado em uma disciplina de testagem psicológica. Um simples olhar para os números da tabela já transmite alguma informação, como o fato de que a maioria dos escores parece ficar entre 30 e 50. Com a estatística descritiva, podemos resumir os dados de modo a facilitar sua compreensão, sendo que uma forma de resumir dados é representá-los graficamente, enquanto outra é condensá-los em estatísticas que representem numericamente a informação em um conjunto de dados.

Distribuições de freqüência

Antes de aplicarmos qualquer fórmula estatística, sempre é uma boa idéia organizar os dados brutos de alguma forma que permita sua inspeção. Normal-

Tabela 2.2 Dados brutos: 60 escores de teste

41	50	39	40	40	31	42	29	37	36
35	45	44	49	38	34	35	32	41	41
39	47	30	45	43	47	35	46	42	41
34	37	38	40	39	39	36	32	48	39
33	42	44	48	47	40	33	46	46	40
44	37	45	43	39	42	37	45	43	38

mente, isso se faz por meio de uma *distribuição de freqüência*. A Tabela 2.3 apresenta uma distribuição dos escores de teste da Tabela 2.2, listando o número de vezes ou a freqüência com que cada escore ocorreu e a percentagem de vezes que aconteceu. A coluna Percentagem Cumulativa mostra a soma consecutiva dos números da coluna Percentagem, do escore mais baixo para o mais alto. Este último conjunto de números nos permite ver a percentagem dos 60 casos que se encaixam em cada escore ou abaixo dele, e, portanto pode ser lido facilmente como escores de postos de percentil.

Tabela 2.3 Distribuição de freqüência de 60 escores de teste

Escores	Freqüência (*f*)	Percentagem[a] (*P*)	Percentagem cumulativa[a] (*PC*)
29	1	1,7	1,7
30	1	1,7	3,3
31	1	1,7	5,0
32	2	3,3	8,3
33	2	3,3	11,7
34	2	3,3	15,0
35	3	5,0	20,0
36	2	3,3	23,3
37	4	6,7	30,0
38	3	5,0	35,0
39	6	10,0	45,0
40	5	8,3	53,3
41	4	6,7	60,0
42	4	6,7	66,7
43	3	5,0	71,7
44	3	5,0	76,7
45	4	6,7	83,3
46	3	5,0	88,3
47	3	5,0	93,3
48	2	3,3	96,7
49	1	1,7	98,3
50	1	1,7	100

[a]Arredondada para a dezena mais próxima

Quando a amplitude de escores é muito grande, *distribuições de freqüência agrupadas* ajudam a organizá-los de forma ainda mais compacta. Nestas distribuições, os escores são agrupados em intervalos de tamanho conveniente para acomodar os dados, e as freqüências são listadas para cada intervalo em vez de para cada um dos escores. Naturalmente, o que se ganha em compacidade é perdido em termos de detalhamento das informações.

Gráficos

Depois de organizados em uma distribuição de freqüência, os dados podem ser transpostos para qualquer um dos diversos formatos gráficos, como gráficos "de pizza" (*pie charts*) ou barras (para dados discretos ou categóricos) e histogramas ou polígonos de freqüência (para dados métricos ou contínuos). Os dados da Tabela 2.3 são mostrados graficamente na forma de um polígono de freqüência na Figura 2.1. Habitualmente, usa-se o eixo horizontal (também denominado *abcissa, linha de base* ou *eixo X*) para representar a amplitude de valores da variável em questão, e o eixo vertical (denominado *ordenada* ou *eixo Y*) para representar as freqüências em que cada valor ocorre na distribuição. As regras e os procedimentos para transformar distribuições de freqüência de vários tipos em gráficos são apresentados na maioria dos livros introdutórios de estatística (ver, p. ex., Kirk, 1999).

Descrição numérica de dados

Além de nos ajudar a visualizar os dados por meio de gráficos, a estatística descritiva também proporciona ferramentas que nos permitem resumir numericamente suas propriedades. Estas ferramentas descrevem a tendência central e a variabilidade dos dados numéricos.

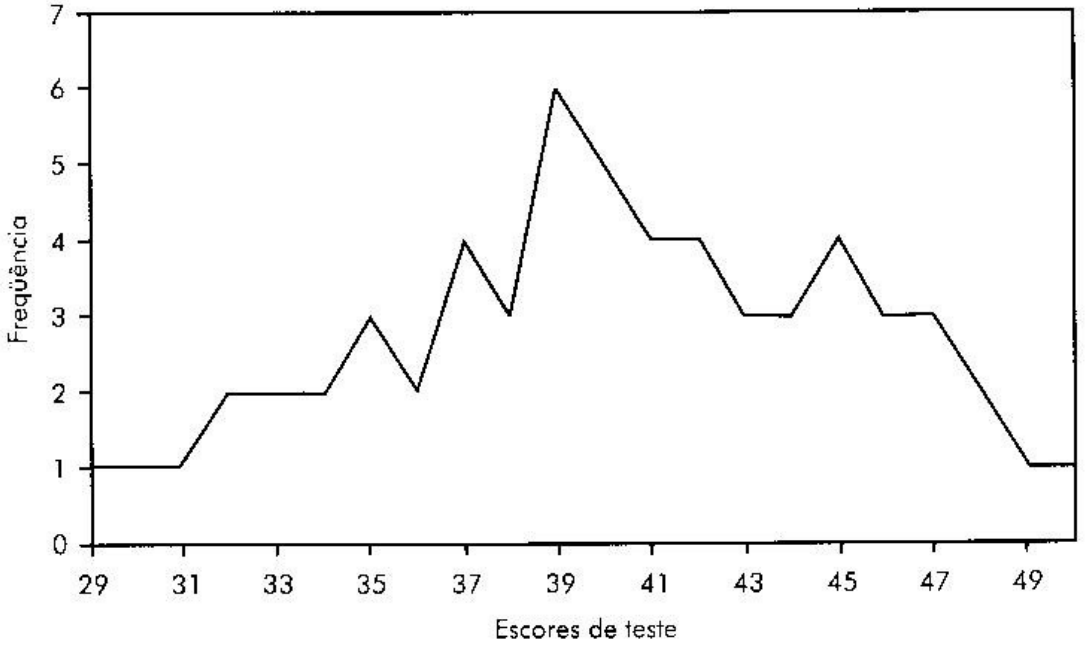

Figura 2.1 Polígono de freqüência para os escores de teste da Tabela 2.3 (n = 60).

Medidas de tendência central

Uma das primeiras coisas a saber ao se inspecionar um conjunto de dados é onde a maior quantidade deles pode ser localizada, bem como seu valor mais representativo ou central. As principais medidas de tendência central – a moda, a mediana e a média – nos informam isso. Assim como qualquer outra estatística, cada uma destas medidas tem vantagens e desvantagens, dependendo do tipo de dados e da distribuição que se quer descrever. Seus méritos e desvantagens relativos, que estão além do âmbito deste livro, também são discutidos na maioria das obras introdutórias de estatística (ver, p. ex., Howell, 2002).

- A *moda,* ou o valor de ocorrência mais freqüente em uma distribuição, é útil primariamente quando lidamos com variáveis qualitativas ou categóricas. Falando estritamente, só pode haver uma moda ou – se não houver variabilidade na distribuição – nenhuma moda. No entanto, se dois ou mais valores de uma distribuição estão ligados à mesma freqüência máxima, a distribuição é denominada *bimodal* ou *multimodal.*
- A *mediana (Mdn)* é o valor que divide em duas metades uma distribuição disposta em ordem de magnitude. Se o número de valores *(n)* da distribuição for ímpar, a mediana é simplesmente o valor do meio; se n for par, a mediana é o ponto médio entre os dois valores do meio.
- A *média* ou média aritmética (μ para uma média populacional e M para uma média de amostra) é obtida somando-se todos os valores de uma distribuição e dividindo o total pelo número de casos da distribuição. Por isso, seu valor efetivo pode estar ou não representado no conjunto de dados. Apesar disso, e do fato de ser a medida de tendência central mais influenciada pelos escores extremos, a média tem muitas propriedades desejáveis que a tornam o indicador de tendência central mais amplamente usado para variáveis quantitativas.

Advertência

Nas páginas a seguir, você vai encontrar algumas fórmulas estatísticas. Caso se sinta tentado a ignorá-las, RESISTA. Lembre-se de que este é um livro sobre fundamentos em testagem psicológica. As únicas fórmulas que você vai encontrar aqui são aquelas que transmitem conceitos essenciais para a compreensão da testagem e do sentido dos escores de teste.

Medidas de variabilidade

Estas estatísticas descrevem quanta *dispersão* existe em um conjunto de dados. Quando somadas a informações a respeito de tendências centrais, as medidas de variabilidade nos ajudam a localizar qualquer valor dentro de uma distribuição e a melhorar a descrição de um conjunto de dados. Embora haja muitas medidas de

variabilidade, os principais índices usados na testagem psicológica são a amplitude, a distância semi-interquartílica, a variância e o desvio padrão.

- A *amplitude* é distância entre dois pontos extremos – os valores mais alto e mais baixo – de uma distribuição. Muito embora seja facilmente calculada, a amplitude é uma medida muito instável porque pode mudar drasticamente devido à presença de um ou dois escores extremos.
- A *distância semi-interquartílica* é a metade da *distância interquartílica* (DIQ), que, por sua vez, é a distância entre os pontos que demarcam o topo do primeiro e do terceiro quartos de uma distribuição. O ponto do primeiro quartil (Q_1), ou 25° percentil, marca o alto do quarto (quartil) mais baixo da distribuição. O ponto do terceiro quartil (Q_3), ou 75° percentil, fica no topo do terceiro quarto da distribuição e marca o início do quartil superior. A distância interquartílica é a amplitude entre Q_1 e Q_3, e, portanto, engloba os 50% que ficam no meio de uma distribuição. No exemplo apresentado na Tabela 2.3, o 25° percentil está no escore de 37, e o 75° percentil está no 44. A distância interquartílica é 44 – 37 = 7, e a distância semi-interquartílica é 7 ÷ 2 = 3,5. Observe que enquanto 53% dos escores se encaixam em uma estreita faixa de 8 pontos, os outros 47% estão dispersos pela amplitude restante de 14 pontos de escore.
- A *variância* é a soma do quadrado das diferenças ou desvios entre cada valor (X) de uma distribuição e a média desta distribuição (M), dividida por N. Mais sucintamente, a variância é a média da *soma dos quadrados (SQ). A soma dos* quadrados é uma abreviação para a soma do quadrado dos valores de desvio ou escores de desvio, $S(X - M)^2$. Os escores de desvio tem que ser elevados ao quadrado antes de serem somados para eliminar números negativos. Se estes números não estiverem ao quadrado, os escores positivos e negativos de desvio em torno da média iriam se cancelar mutuamente e sua soma seria zero. A soma dos quadrados representa a quantidade total de variabilidade em uma distribuição de escores, e a variância (*SQ*/N) representa sua variabilidade média. Devido à elevação ao quadrado dos escores de desvio, no entanto, a variância não é expressa nas mesmas unidades que a distribuição original.
- O *desvio padrão* é a raiz quadrada da variância. Juntamente com esta, proporciona um único valor que é representativo das diferenças individuais ou desvios em um conjunto de dados – calculados a partir de um ponto de referência comum, qual seja, a média. O desvio padrão é uma medida da variabilidade média de um conjunto de escores, expresso nas mesmas unidades que estes. É a medida primordial de variabilidade para a testagem, bem como para muitos outros fins, e é útil em diversas manipulações estatísticas.

O quadro Consulta Rápida 2.3 lista alguns dos símbolos básicos de notação que serão usados neste livro, juntamente com as fórmulas para média, distância interquartílica, desvio padrão e variância. As medidas de tendência central e variabilidade para os 60 escores de teste da Tabela 2.3 são listadas na Tabela 2.4. Embo-

CONSULTA RÁPIDA 2.3

Notação básica e fórmulas

X = Um dado ou valor em uma distribuição; em testagem psicológica, X quase sempre representa um escore bruto.
S = Soma de.
n = Tamanho da amostra, ou seja, o número total de casos de uma distribuição; em testagem psicológica, n quase sempre representa o número de pessoas ou de escores.
N = Tamanho da população

M, ou $\bar{X}$ = Média de $X = \frac{\Sigma X}{n}$

μ = Média populacional
Mdn = Mediana = 50º percentil.
Q_1 = ponto do 1º quartil = 25º percentil.
Q_3 = ponto do 3º quartil = 75º percentil.
$Q_1 - Q_3$ = Distância interquartílica (DIQ).
DIQ ÷ 2 = distância semi-interquartílica.

s^2 = Variância da amostra = $\frac{\Sigma(X - M)^2}{n - 1}$

σ^2 = Variância populacional = $\frac{\Sigma(X - \mu)^2}{N}$

d ou DP = Desvio padrão da amostra = $\sqrt{s^2}$

σ = Desvio padrão populacional = $\sqrt{\sigma^2}$

Tabela 2.4 Estatística descritiva para os 60 escores de teste da Tabela 2.3

Medidas de tendência central		Medidas de variabilidade	
Média	= 40,13	Amplitude = 50 – 29	= 21
Mediana	= 40,00	Variância	= 25,745
Moda	= 39	Desvio padrão	= 5,074
Q1 ou 25º percentil	= 37	Distância interquartílica = Q3 - Q1 = 44 – 37	= 7
Q3 ou 75º percentil	= 44	Distância semi-interquartílica = 7 ÷ 2	= 3,5

ra informações detalhadas sobre os 60 escores não estejam disponíveis, as estatísticas da Tabela 2.4 descrevem concisamente onde os escores se agrupam e a quantidade média de dispersão do conjunto de dados.

A importância da variabilidade

Embora possa ser verdade que a variedade é o tempero da vida, ela é o ingrediente principal da testagem psicológica, que depende da variabilidade entre os indivíduos. Sem diferenças individuais, não haveria variabilidade, e os testes não nos ajudariam a fazer determinações ou tomar decisões a respeito de pessoas. Se todos os outros fatores forem iguais, quanto maior a quantidade de variabilidade entre os indivíduos em termos da característica que estamos tentando avaliar, mais precisamente poderemos fazer distinções entre eles. Conhecer os formatos da distribuição de escores, bem como suas tendências centrais e variabilidades proporciona a base para boa parte dos fundamentos em interpretação de escores de teste discutidos no Capítulo 3.

Pondo em prática

- Vá à Tabela 2.3 e conte quantos escores estão dentro de ± 1 DP da média – isto é, entre 40 ± 5.
- Verifica-se que 41 dos 60 escores, aproximadamente 2/3 deles, se situam entre 35 e 45.
- Esta proporção é típica das distribuições que se aproximam do formato da curva normal, como a distribuição da Figura 2.1.

O MODELO DA CURVA NORMAL

Definição

A curva normal, também conhecida como *curva do sino*, é uma distribuição em alguns aspectos semelhante à da Figura 2.1. Sua linha de base, equivalente ao eixo X da distribuição da Figura 2.1, mostra as unidades de desvio padrão (σ); o eixo vertical, ou ordenada, geralmente não precisa ser mostrado porque a curva normal não é uma distribuição de freqüência de dados, mas um modelo matemático de uma distribuição ideal ou teórica. A altura que a curva alcança em cada ponto ao longo da linha de base é determinada por uma fórmula matemática que descreve as relações específicas a partir do modelo e estabelece a forma e as proporções exatas da curva. Como todos os modelos ideais, a curva normal não existe, baseia-se na teoria da probabilidade. Felizmente, para nossos propósitos, podemos compreender os fatos básicos relativos à curva normal sem sabermos muito a respeito de suas bases matemáticas.

Embora o modelo da curva normal seja um ideal, a distribuição de dados reais muitas vezes se aproxima dela, como é o caso dos dados da Tabela 2.3 apresentados na Figura 2.1. A semelhança entre o modelo e as distribuições de muitas variáveis no mundo natural tornou-o útil na estatística descritiva. Ainda mais importante é o fato de que muitos eventos do acaso, se repetidos por um número suficientemente grande de vezes, geram distribuições que se aproximam da curva

normal. É esta ligação com a teoria da probabilidade que faz com que a curva normal desempenhe um papel importante na estatística inferencial. Como veremos a seguir, a utilidade do modelo da curva normal deriva de suas propriedades.

Propriedades do modelo da curva normal

Muitas propriedades do modelo da curva normal são claramente evidentes com a simples inspeção visual (ver Figura 2.2). Por exemplo, pode-se ver que a distribuição normal tem as seguintes propriedades:

- Tem *formato de sino*, como indica seu "apelido".
- É *bilateralmente simétrica*, o que significa que suas duas metades são idênticas (se dividirmos a curva em duas partes, cada metade contém 50% da área sob a curva).
- Tem caudas que se aproximam da linha de base mas nunca a tocam, e por isso seus limites se estendem ± ao infinito (± ∞), uma propriedade que explicita a natureza teórica e matemática da curva.
- É unimodal, ou seja, tem um único ponto de freqüência máxima ou altura máxima.
- Tem *média*, *mediana* e *moda* que coincidem no centro da distribuição, porque o ponto onde a curva está em equilíbrio perfeito, que é a média, também é o ponto que divide a curva em duas metades iguais, que é a mediana, e é o valor mais freqüente, que é a moda.

Além destas propriedades, a curva normal tem outras características menos óbvias que estão ligadas à sua regra de função matemática. Esta fórmula – que não é essencial – está disponível na maioria dos livros de estatística e em algumas das

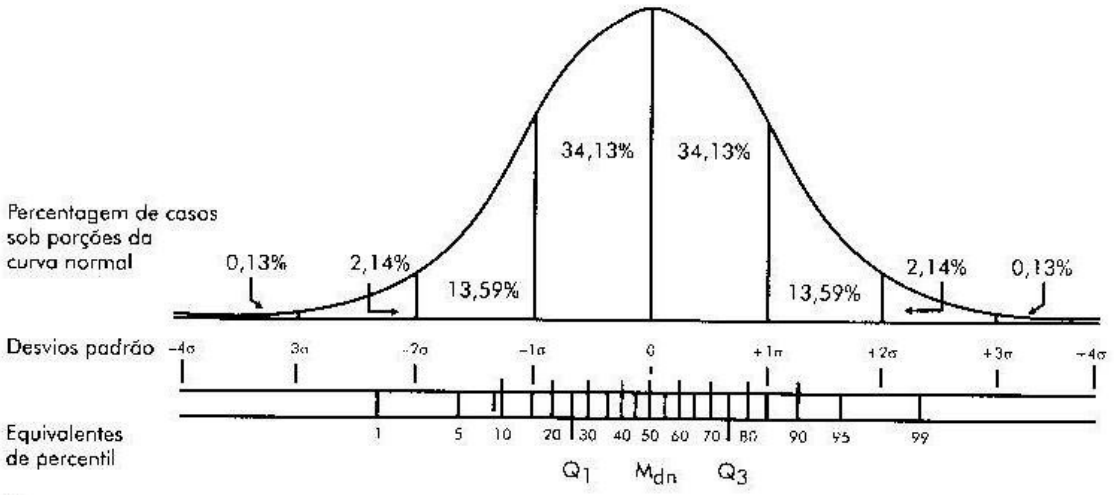

Figura 2.2 A curva normal, com percentagens de casos em cada segmento de unidade de σ de −4 a +4, percentagens cumulativas e equivalentes de percentil.

páginas da Internet sobre a curva normal mencionadas no quadro Consulta Rápida 2.4. Ela envolve dois elementos constantes (μ e e) e dois valores que podem variar. Cada curva normal em particular é apenas um exemplo de uma família de distribuições de curva normal que difere em função de seus dois valores variáveis. Estes são a média, designada como μ, e o desvio padrão, designado como s. Depois que os parâmetros de μ e σ para uma distribuição normal são determinados, pode-se calcular a altura da ordenada (eixo Y), em cada ponto ao longo da linha de base (eixo X) com a fórmula que define a curva. Quando a curva normal tem média zero e desvio padrão 1, é denominada *distribuição normal padrão*. Uma vez que a área total sob a curva normal equivale à unidade (1,00), o conhecimento da altura da curva (a ordenada Y) em qualquer ponto ao longo da linha de base, ou eixo X, nos permite calcular a proporção *(p)* ou percentagem *(p* x 100) da área sob a curva que está acima e abaixo de qualquer valor de X, bem como entre quaisquer dois valores de X. A tabela estatística que resulta destes cálculos, que mostra as áreas e as ordenadas da curva normal padrão, está disponível no Apêndice C, juntamente com uma explicação básica de seu uso.

Na curva normal, as unidades de desvio padrão ou σ são posicionadas em distâncias iguais ao longo do eixo X, em pontos que marcam as inflexões da curva (isto é, os pontos em que a curva muda de direção). A Figura 2.2 mostra a curva normal dividida em cada unidade σ de –4 a +4, bem como as percentagens da área contidas em cada segmento. Observe que se somarmos todas as percentagens das áreas acima da média, o resultado equivale a 50%, assim como a soma de todas as áreas abaixo da média. Além disso, a área entre $+1\sigma$ e -1σ é de 68,26% (34.13% x 2) – aproximadamente 2/3 da curva –, e a área entre $+2\sigma$ e -2σ é de 95,44%, quase que a curva inteira. O conhecimento desses fatos básicos a respeito da curva normal é extremamente útil em estatística.

CONSULTA RÁPIDA 2.4

Páginas da internet sobre a curva normal

Estas são apenas três das muitas páginas da Internet que podem ser encontradas digitando-se "a curva normal" em um bom mecanismo de busca *on-line*:

- http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html
  Esta página traz uma demonstração simples e visualmente atraente do processo que resulta na curva do sino.
- http://stat-www.berkeley.edu/~stark/Java/NormHiLite.htm
  Esta página tem uma ferramenta interativa que permite destacar qualquer segmento da curva normal e ver imediatamente qual percentagem na área está contida no segmento destacado.
- http://www.psychstat.smsu.edu/introbook/sbk11.htm
  Esta é uma das muitas páginas que explica fatos básicos a respeito da curva normal de forma clara e sucinta.

USOS DO MODELO DA CURVA NORMAL

Usos descritivos

Uma vez que as proporções da área sob a curva normal padrão que se situam acima e abaixo de qualquer ponto da linha de base ou entre quaisquer dois pontos desta são preestabelecidas – e fáceis de achar nas tabelas de áreas da curva normal como a que é apresentada no Apêndice C – podemos aplicar prontamente estas proporções a qualquer outra distribuição que tenha forma semelhante. Na testagem, esta particular aplicação da distribuição normal é usada repetidamente na geração dos escores padrões descritos no próximo capítulo.

Em algumas circunstâncias, mesmo quando uma distribuição se aproxima, mas não equivale exatamente à curva normal, ainda podemos usar as proporções do modelo da curva normal para regularizar os escores. A *normalização* de escores envolve transformá-los de tal forma que eles tenham o mesmo sentido, em termos de sua posição, que teriam se pertencessem a uma distribuição normal. Este procedimento, que não é tão complicado quanto pode parecer, faz uso das percentagens cumulativas calculadas de uma distribuição de freqüência (ver Tabela 2.3) e será discutido mais detalhadamente e com exemplos no capítulo seguinte.

Usos inferenciais do modelo da curva normal

Na estatística inferencial, o modelo da curva normal é útil para: (a) estimar parâmetros populacionais e (b) testar hipóteses a respeito de diferenças. As aplicações do modelo da curva normal à estimativa de parâmetros populacionais e à testagem de hipóteses faz uso de duas noções inter-relacionadas, quais sejam, distribuições de amostragem e erros padrões.

Distribuições de amostragem são distribuições hipotéticas, e não reais, de valores bascadas na premissa de que um número infinito de amostras de um determinado tamanho podem ser derivadas de uma população. Se isso fosse feito, e se as estatísticas destas amostras fossem registradas, muitas (mas não todas) distribuições resultantes das estatísticas ou distribuições de amostragem seriam normais. A média de cada distribuição hipotética de amostragem seria igual ao parâmetro populacional e o desvio padrão da distribuição de amostragem seria o erro padrão das estatísticas em questão.

O *erro padrão (EP)* de uma estatística obtida de uma amostra é, portanto, concebido como o desvio padrão da distribuição de amostragem que resultaria se obtivéssemos a mesma estatística de um grande número de amostras de tamanho igual, determinadas aleatoriamente. Pode ser facilmente calculado usando estatísticas da amostra (ver, p. ex., a Tabela 2.5). Depois de obtermos uma determinada estatística de uma amostra e seu erro padrão, a premissa de uma distribuição normal de amostragem nos permite usar as áreas da curva normal para estimar os parâmetros populacionais baseados na estatística obtida.

Tabela 2.5 Erro padrão da média para os dados das Tabelas 2.3 e 2.4

Média (M) = 40,13 Desvio padrão (s) = 5,074 Tamanho da amostra (n) = 60

Erro padrão da média (SE_M)* $= \frac{s}{\sqrt{n}} = \frac{5,074}{\sqrt{60}} = \frac{5,074}{7,7459} = 0,655$

— Não esqueça

O Apêndice C contém a Tabela de Áreas e Ordenadas da Curva Normal, juntamente com uma explicação de como ela é usada. Como se aplica a todas as fórmulas que constam deste livro, as informações do Apêndice C são apresentadas unicamente porque a familiaridade com as mesmas é um requisito essencial para a compreensão de escores de teste.

Estimativa de parâmetros populacionais

Um exemplo hipotético

Para estimar um parâmetro populacional, como a altura média de uma mulher adulta dos Estados Unidos, podemos obter uma amostra aleatória de 50 mulheres adultas, uma de cada estado do país. Podemos supor que a altura média para esta amostra, que seria uma *estimativa* da média populacional, é de 64 polegadas, e também que o desvio padrão é de 4 polegadas. Se repetíssemos este procedimento infinitas vezes, obtendo um número infinito de amostras de 50 mulheres cada uma e registrando as médias de todas estas amostras, a distribuição de amostragem das médias resultante corresponderia ao modelo da curva normal. A média desta distribuição teórica de amostragem pode ser entendida como a média populacional (isto é, a altura média de todas as mulheres adultas dos Estados Unidos).

Obviamente, este curso de ação não é apenas pouco prático, mas também impossível. Por isso, usamos a estatística inferencial para estimar a média populacional. Encontramos erro padrão da média (SE_M) com a fórmula $s/\sqrt{n}$, em que s é o desvio padrão da amostra (4) e n é o número de casos da amostra (50). Neste caso, 4 dividido pela raiz quadrada de 50 é igual a 7,07, o que produz um $SE_M = 0,565$. Assim, baseados nas estatísticas obtidas da amostra, podemos dizer que a altura média das mulheres adultas dos Estados Unidos está dentro da faixa de nossa média obtida de 64 polegadas ± 0,565 polegadas, ou entre 63,435 polegadas e 64,565 polegadas. Somar e subtrair 1 SE_M da média da amostra nos dá um intervalo de confiança de 68% para a média populacional, porque 68% da área sob a curva normal se encaixa dentro de $\pm 1\sigma$ (ou, neste caso, ± 1 SE_M). Se quisermos

*N. de R.T. A sigla será utilizada em inglês.

fazer uma afirmação com nível mais alto de confiança, podemos escolher um intervalo de confiança maior selecionando um número maior de unidades de σ e multiplicando-o pelo SE_M. Como vemos na Figura 2.2, o segmento entre ±2σ engloba 95,44% da área sob a curva normal; portanto, em nosso exemplo, o intervalo entre 64±2 SE_M ou 64±2 (0,565 pol.) = 64±1,13 polegadas e engloba a amplitude de 62,87 a 65,13 polegadas, dentro da qual podemos ter confiabilidade de 95,44% de que a altura média das mulheres adultas dos Estados Unidos aí se localiza.

Exemplo com dados da Tabela 2.3: Se calcularmos o erro padrão da média (SE_M) para os dados da Tabela 2.3 usando a fórmula $s/\sqrt{n}$, na qual s é o desvio padrão (5,074) e n é o número de casos da amostra (60), o SE_M = 0,655 (ver Tabela 2.5). Se a amostra de 60 estudantes tivesse sido escolhida aleatoriamente entre todos os estudantes que já se submeteram àquele teste em particular, poderíamos então pressupor que existe uma probabilidade de aproximadamente 68% de que a média da população de *todos* os estudantes que se submeteram ao teste está dentro de ± 0,655 pontos, ou ±1 SE_M, da média obtida de 40,13, ou em algum ponto da amplitude de 39,475 a 40,785. Da mesma forma, podemos dizer com confiança de 95,44% – o que significa que nossas chances de estarmos errados são de menos de 5% – que o intervalo entre a média de 40,13 ±2 SE_M, isto é, a amplitude de 38,82 a 41,44, inclui a média populacional.

A significância dos erros padrões

Os erros padrões são extremamente importantes na estatística inferencial. Em ambos os exemplos apresentados, podemos estimar as amplitudes em que os parâmetros populacionais podem ser encontrados a partir das premissas de que: (a) a média obtida da mostra é a melhor estimativa que temos da média populacional e (b) o erro padrão da média é equivalente ao desvio padrão da distribuição de amostragem hipotética das médias, a qual se supõe que seja normal. Premissas semelhantes, juntamente com as estimativas fornecidas pelas áreas sob a curva normal padrão e outras distribuições teóricas que aparecem em tabelas estatísticas – como a distribuição *t* de Student – podem ser usadas não apenas para gerar afirmações de probabilidade a respeito de parâmetros populacionais derivados de outras estatísticas de amostras, mas também para gerar afirmações de probabilidade a respeito das diferenças obtidas entre estatísticas de amostras.

Quando se testa a significância das diferenças entre as médias ou proporções de amostras, as diferenças obtidas são divididas pelos erros padrões destas diferenças, calculados por fórmulas apropriadas para o tipo específico de diferença a ser testada. As razões resultantes, chamadas *razões críticas*, juntamente com as distribuições apropriadas para a estatística em questão, podem então ser usadas para determinar a probabilidade de que uma diferença obtida possa ter resultado do acaso. Embora a maioria das técnicas de estatística inferencial estejam muito além do âmbito deste livro, vamos abordar os erros padrões novamente em conexão com a fidedignidade e a validade de escores de teste nos Capítulos 4 e 5. O quadro Consulta Rápida 2.5 resume as principais razões por que o modelo da curva normal é tão importante no campo da testagem psicológica.

— Não esqueça

Dois conceitos essenciais da estatística inferencial: distribuições de amostragem e erros padrões

- *Distribuições de amostragem* são distribuições teóricas dos valores de uma variável, ou de uma estatística, que resultariam da coleta e registro de valores (p. ex., escores) ou estatísticas (p. ex., médias, desvios padrões, coeficientes de correlação, etc.) de um número infinito de amostras de um determinado tamanho de uma população em particular. As distribuições de amostragem não existem na realidade: são constructos hipotéticos usados para determinar estimativas de probabilidade de valores ou estatísticas obtidos através de um artifício conhecido como *erro padrão*.
- Os *erros padrões* são entidades estatísticas que podem ser calculadas através de várias fórmulas a partir de dados de amostras; fornecem os meios para compararmos valores ou estatísticas obtidos de amostras de suas distribuições de amostragem teóricas. Um erro padrão é o desvio padrão estimado da distribuição de amostragem teórica de um valor ou estatística obtidos.

CONSULTA RÁPIDA 2.5

Por que a curva normal é tão importante na testagem psicológica?

No testagem, o modelo da curva normal é usado de modos paralelos à distinção entre estatística descritiva e inferencial:

1. O modelo da curva normal é usado *descritivamente* para localizar a posição de escores derivados de distribuições normais. Em um processo conhecido como *normalização*, descrito no Capítulo 3, a curva normal também é usada para se fazer distribuições que não são normais – mas se aproximam do normal – conforme o modelo, em termos das posições relativas dos escores.
2. O modelo da curva normal se aplica *inferencialmente* nas áreas de: (a) *fidedignidade*, para derivar intervalos de confiança que avaliem escores obtidos e as diferenças entre eles (ver Capítulo 4), e (b) *validade*, para derivar intervalos de confiança para predições ou estimativas baseadas em escores de testes (ver Capítulo 5).

Distribuições não-normais

As representações gráficas de distribuições obtidas permitem a comparação das distribuições de freqüência com a distribuição normal. Isto é muito importante porque, na medida em que um polígono de freqüência ou histograma difere em formato da curva normal, as proporções da área sob a curva não mais se aplicam. Além disso, o modo particular como as distribuições diferem da curva normal pode ter implicações significativas para os dados.

Existem muitas diferenças possíveis entre as distribuições obtidas e o modelo da curva normal. O modo como aquelas se desviam e o grau em que o fazem têm implicações para a quantidade de informação que as distribuições transmitem. Um caso extremo pode ser ilustrado pela distribuição resultante se todos os valores de um conjunto de dados ocorressem com a mesma freqüência. Esta distribuição, que teria formato retangular, não implicaria qualquer diferença na probabilidade de ocorrência de determinado valor, e por isso não seria útil na tomada de decisões a respeito do que estivesse sendo medido.

Um tipo diferente e mais plausível de desvio do modelo da curva normal acontece quando as distribuições têm duas ou mais modas. Se uma distribuição for bimodal ou multimodal, é necessário considerar a possibilidade de problemas de amostragem ou características especiais da mostra. Por exemplo, uma distribuição de notas em uma turma de alunos na qual as freqüências máximas ocorrem nas notas A e D, com muito poucas notas B ou C, pode significar que estes alunos são atípicos em algum aspecto ou que pertencem a grupos com diferenças significativas de preparação, motivação ou nível de habilidade. É óbvio que informações dessa natureza quase invariavelmente teriam implicações importantes; no caso deste exemplo, poderiam levar a professora a dividir a turma e usar diferentes abordagens pedagógicas com cada grupo.

Duas outras formas de desvio do modelo da curva normal têm implicações significativas para dados de testes, e dizem respeito às propriedades da curtose e da assimetria (*skewness*) das distribuições de freqüência.

Curtose

Este termo, bastante estranho, que deriva da palavra grega para *convexidade*, simplesmente se refere ao aspecto achatado ou pontiagudo de uma distribuição. A curtose está diretamente relacionada à quantidade de dispersão de uma distribuição. As distribuições *platicúrticas* têm maior quantidade de dispersão, demonstrada por caudas mais extensas, e as distribuições *leptocúrticas* têm quantidades menores. A distribuição normal é *mesocúrtica*, o que significa que ela tem um grau intermediário de dispersão.

A curtose aplicada: A hipótese da maior variabilidade masculina. No campo da psicologia diferencial, uma hipótese antiga afirma que a variação de inteligência é maior entre os homens do que entre as mulheres. Esta hipótese surgiu da observação de uma super-representação dos homens entre pessoas com realizações extraordinárias e outras em instituições para retardados mentais. Embora tenha havido muita discussão e um respaldo moderado a esta hipótese (ver, p. ex., Halpern, 1997; Hedges e Nowell, 1995), por uma variedade de razões – incluindo a natureza dos testes de inteligência – a questão ainda não foi resolvida. Se a hipótese da maior variabilidade masculina se mostrar verdadeira, isto significa que mais homens do que mulheres estão localizados nas extremidades alta e baixa da distribuição dos escores de testes de inteligência. Neste caso, as distribuições para mulheres e homens iriam diferir em curtose. Suas representações gráficas, caso fossem sobrepostas, poderiam ter a aparência das distribuições hipotéticas da Figura 2.3, que mostram uma distribuição leptocúrtica para as mulheres e uma distribuição platicúrtica para os homens, sem diferenças nos escores médios entre os dois gêneros.

Assimetria

O termo em inglês *skewness* (*Sk*) expressa a falta de simetria. Como vimos, a distribuição normal é perfeitamente simétrica, com $Sk = 0$, maior volume no meio e

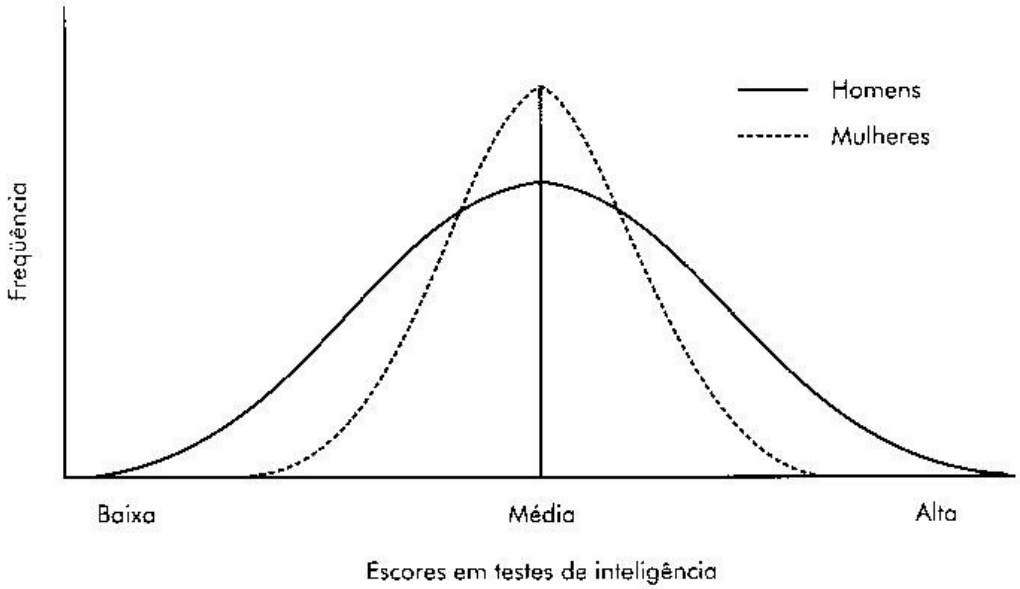

Figura 2.3 Distribuições hipotéticas de escores em testes de inteligência, mostrando maior variabilidade masculina que feminina (curva platicúrtica *versus* leptocúrtica).

duas metades idênticas. Uma distribuição também pode ser assimétrica. Se a maior parte dos valores está na extremidade superior da escala e a cauda mais longa se estende na direção da extremidade inferior, a distribuição é *negativamente* assimétrica ($Sk < 0$). Por outro lado, se a maior parte dos valores estiver na parte inferior e a cauda mais longa se estender na direção do alto da escala, a distribuição será *positivamente* assimétrica ($Sk > 0$).

A assimetria aplicada. O significado da assimetria em relação às distribuições de escores de teste é fácil de ser identificado. Se uma distribuição é negativamente assimétrica, isso significa que a maioria das pessoas obteve escores altos; se for positivamente assimétrica, significa que a maior parte das pessoas teve escores baixos. A Figura 2.4 mostra exemplos de distribuições positiva e negativamente assimétricas. O Painel A da figura mostra uma distribuição positivamente assimétrica de escores em um teste no qual a maioria dos estudantes teve pontuação baixa, e o Painel B mostra uma distribuição negativamente assimétrica de escores em um teste no qual a maioria dos testandos teve pontuação alta.

Por que a forma das distribuições é relevante para a testagem psicológica?

Quando um teste está sendo desenvolvido, a forma e as características das distribuições de escore obtidas com suas versões preliminares ajudam a determinar os ajustes necessários. O formato das distribuições de escore obtidas durante o processo de desenvolvimento do teste deve corresponder às expectativas baseadas no que está sendo medido e no tipo de testandos incluídos nas amostras preliminares

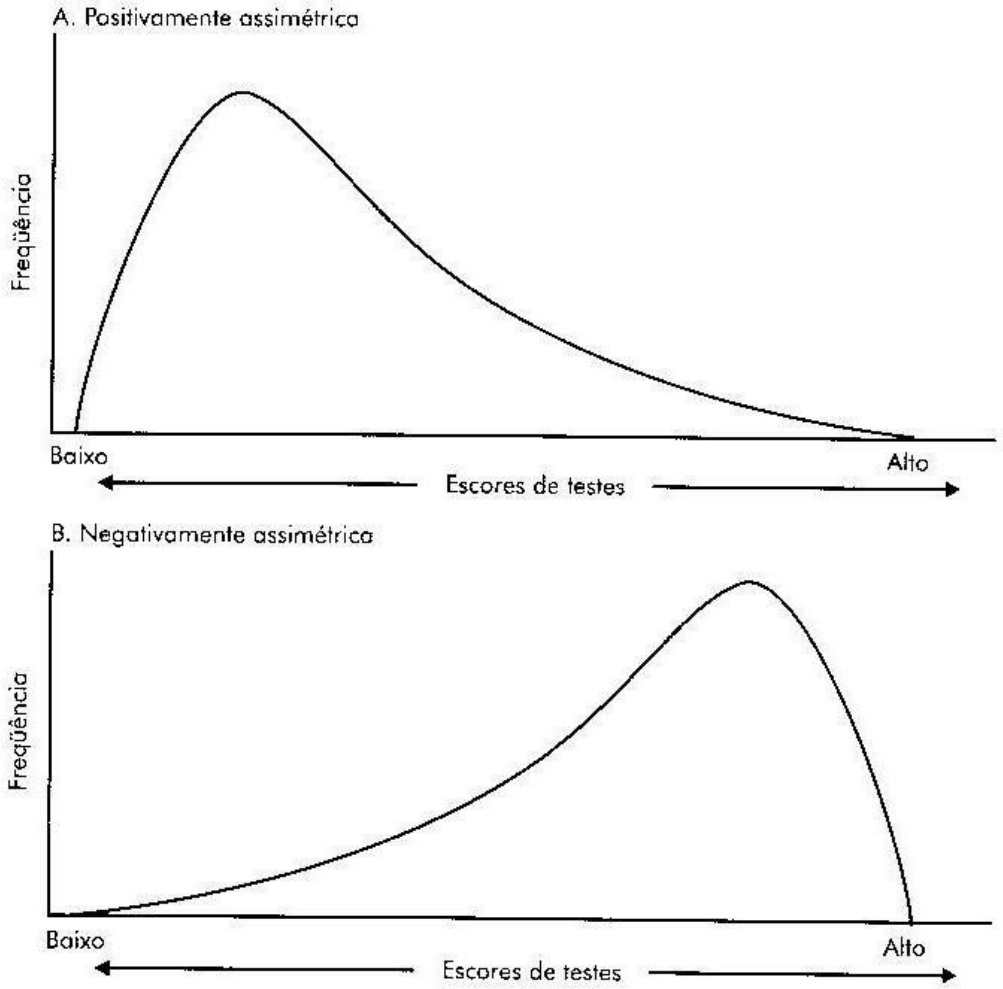

Figura 2.4 Distribuições assimétricas.

ou de padronização. Por exemplo, se um teste de realização é voltado para estudantes universitários e a distribuição dos escores de uma amostra representativa desta população for negativamente assimétrica, isso significa que o teste pode ser fácil demais, e seu criador pode ter que acrescentar itens mais difíceis para deslocar a maior parte dos escores na direção do centro da distribuição. Inversamente, se o mesmo teste for aplicado a uma amostra representativa de estudantes do ensino fundamental e a distribuição de seus escores for positivamente assimétrica, o resultado estará de acordo com as expectativas e não haverá necessidade de ajustes.

FUNDAMENTOS DE CORRELAÇÃO E REGRESSÃO

Até agora, nossa discussão se concentrou, primariamente, na descrição e no tratamento de estatísticas derivadas de mensurações de uma única variável, ou estatís-

ticas *univariadas*. Se estivéssemos interessados apenas em escores de testes (uma possibilidade improvável), estas estatísticas seriam suficientes. No entanto, para que tenham algum sentido na arena prática, os escores de testes precisam oferecer informações a respeito de outras variáveis que são significativas no mundo real. Os métodos correlacionais são as técnicas usadas para obter índices do grau em que duas ou mais variáveis estão relacionadas mutuamente, índices que são chamados de *coeficientes de correlação*. Os métodos correlacionais são as principais ferramentas que temos para demonstrar ligações: (a) entre escores em testes diferentes; (b) entre escores de teste e variáveis que não pertencem a testes; (c) entre escores em partes de um teste ou itens de um teste e o escore no teste inteiro; (d) entre escores parciais ou escores em itens e variáveis que não pertencem ao teste e (e) entre escores em diferentes partes de um teste ou diferentes itens de um único teste. Devido a estas múltiplas aplicações, a noção de correlação desempenha um papel importante nas discussões a respeito de fidedignidade, validade e desenvolvimento de testes que veremos nos próximos capítulos.

Com a correlação, entramos no campo das estatísticas *bivariadas* ou *multivariadas*. Em vez de termos uma única distribuição de freqüência de medidas em uma variável, precisamos de pelo menos dois conjuntos de medidas ou observações do mesmo grupo de pessoas (p. ex., escores em dois testes diferentes) ou pares combinados de observações para dois conjuntos de indivíduos (p. ex., os escores de pares de gêmeos no mesmo teste). Quando calculamos um coeficiente de correlação para descrever a relação entre duas variáveis, os dados são organizados na forma de distribuições bivariadas, como as apresentadas na Tabela 2.6 para dois conjuntos de dados fictícios.

Para calcular um coeficiente de correlação, tudo o que precisamos são os dados (isto é, observações) de duas variáveis. Estas podem ser a renda anual e os anos de escolaridade para um conjunto de pessoas, a quantidade de chuva e o tamanho das colheitas para um período de vários anos, o comprimento médio das saias femininas e o desempenho do mercado de ações ao longo de um período de tempo, a posição do sol no céu e a quantidade de luz em um determinado local, os escores em um teste e um índice de desempenho no trabalho para um grupo de empregados, etc. Duas variáveis (geralmente designadas X e Y) podem ser correlacionadas usando-se qualquer um de vários métodos correlacionais que diferem em termos dos tipos de dados e de relações para os quais são apropriados.

Correlação linear

A relação entre duas variáveis é dita *linear* quando a direção e a taxa de mudança de uma variável são constantes em relação às mudanças da outra. Quando dispostos em um gráfico, os pontos de dados para este tipo de relação formam um padrão elíptico reto ou quase reto. Se houver uma correlação entre duas variáveis e a relação entre elas for linear, existem apenas dois resultados possíveis: (a) uma correlação positiva ou (b) uma correlação negativa. Se não houver correlação, os pontos de dados não se alinham em qualquer padrão ou tendência definida, e

Tabela 2.6 Dois conjuntos de dados bivariados

Indivíduo	Escore no teste X	Escore no teste Y
A. Dados para uma correlação positiva perfeita		
1	3	5
2	4	6
3	6	8
4	7	9
5	8	10
6	9	11
7	10	12
8	11	13
9	13	15
10	14	16
B. Dados para uma correlação negativa perfeita		
1	140	5
2	130	6
3	110	8
4	100	9
5	90	10
6	80	11
7	70	12
8	60	13
9	40	15
10	30	16

podemos pressupor que os dois conjuntos de dados não compartilham uma fonte comum de variância. Se houver uma correlação positiva ou negativa de qualquer magnitude, podemos avaliar a possibilidade de que a correlação tenha resultado do acaso, usando o tamanho da amostra na qual a correlação foi calculada e tabelas estatísticas que mostram a probabilidade de ocorrência ao acaso de um coeficiente de uma determinada magnitude. Naturalmente, quanto maior o coeficiente, menor a probabilidade de que possa ser resultado do acaso. Se a probabilidade de que o coeficiente obtido tenha resultado no acaso for muito pequena, podemos ter confiança de que a correlação entre X e Y é maior do que zero. Nestes casos, pressupomos que as duas variáveis compartilham uma certa quantidade de variância comum. Quanto maior e mais estatisticamente significativo for um coeficiente de correlação, maior a quantidade de variância que podemos pressupor entre X e Y. A proporção de variância compartilhada por duas variáveis muitas vezes é estimada elevando-se ao quadrado o coeficiente de correlação (r_{xy}) e obtendo o *coeficiente de determinação*, ou r^2_{xy}. Embora os coeficientes de determinação nos informem quanto da variância de Y pode ser explicada pela variância de X, ou vice-versa, eles não necessariamente indicam que existe uma relação causal entre X e Y.

Gráficos de dispersão

A representação gráfica de dados bivariados na forma de diagramas ou gráficos de dispersão é essencial para visualizarmos o tipo de relação do qual estamos tratando. Os gráficos de dispersão da Figura 2.5 apresentam os padrões de pontos que resultam da representação das distribuições bivariadas da Tabela 2.6. Estas figuras nos permitem literalmente ver a força e a direção da relação entre as duas variáveis de cada conjunto. Podemos ver que em ambas as partes da figura os padrões for-

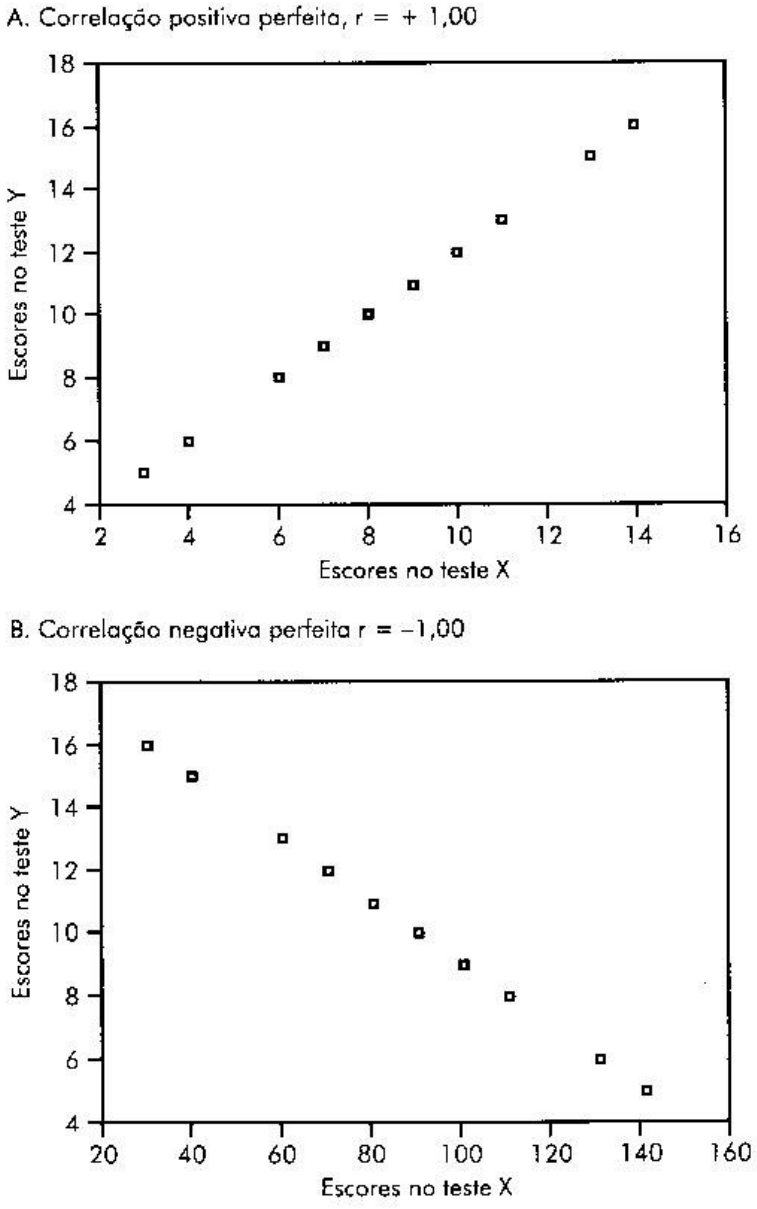

Figura 2.5 Gráficos de dispersão dos dados bivariados da Tabela 2.6.

mam uma linha diagonal reta, indicando que ambas as relações são lineares e fortes (na verdade, estas correlações são perfeitas, o que raramente se vê com dados reais). Uma correlação forte significa que a medida em que os valores de uma variável aumentam ou diminuem existe uma quantidade correspondente de mudança nos valores da outra variável. A direção do padrão de pontos em um gráfico de dispersão indica se as mudanças correspondentes ocorrem na mesma direção ou em direções opostas. No padrão perfeito mostrado na Figura 2.5, Painel A, a relação é invariante: para cada unidade de aumento nos escores do teste X existe um aumento correspondente de uma unidade nos escores do teste Y. Na Figura 2.5, Painel B, vemos outro padrão perfeito invariante: para cada diminuição de 10 unidades no teste X existe um aumento correspondente de uma unidade no teste Y. As relações estão em direções opostas, mas ambas mantêm uma correspondência perfeita em relação a suas respectivas escalas.

A descoberta da regressão

O leitor recorda do Capítulo 1 que Francis Galton fez contribuições significativas para o desenvolvimento da psicometria. Uma das mais importantes foi a descoberta do fenômeno que Galton denominou *regressão*, uma descoberta que resultou de suas tentativas de mapear a semelhança entre pais e filhos em diversas variáveis e produzir evidências de sua natureza hereditária. Em termos da variável da altura, por exemplo, Galton descobriu que os pais que eram mais altos do que a média tendiam a ter filhos que, quando adultos, também eram mais altos do que a média dos pais da amostra, mas mais próximos desta média do que os próprios pais. O inverso também se aplicava aos pais que eram mais baixos do que a média: seus filhos, quando adultos, também tendiam a ser mais baixos do que a média dos pais da amostra de Galton, mas mais próximos desta média do que seus próprios pais. Quando dispôs em gráfico estes dados bivariados de conjuntos pareados de alturas de pais e filhos, bem como outros conjuntos de variáveis, Galton percebeu que este padrão de *regressão em direção à média* continuava a se repetir, ou seja, escores extremos dos pais em uma variável tendiam a estar associados a escores mais próximos da média nos filhos. Além disso, Galton constatou que, se representasse graficamente a altura dos filhos em várias faixas, relativas às alturas médias dos pais dentro daqueles intervalos, ele obteria um padrão linear, o qual denominou *linha de regressão*. Galton compreendeu que a inclinação da linha de regressão representava a força ou magnitude da relação entre as alturas de pais e filhos: quanto maior a inclinação da linha, mais forte a relação entre as duas variáveis.

Apesar da significância de sua descoberta, as conclusões de Galton a respeito do fenômeno da regressão não foram muito precisas (ver Cowles, 2001). Isso foi, em parte, resultado de restrições nos dados que ele usou em suas análises e, em parte, devido à sua interpretação equivocada das causas das correlações entre variáveis. Como as bases genéticas da hereditariedade não estavam claras na época em que Galton se dedicou a estes problemas, seus equívocos na interpretação da regressão são compreensíveis. Mesmo assim, os procedimentos que ele desenvolveu para retratar a relação entre duas variáveis provaram ser extremamente úteis na

avaliação da quantidade de variância compartilhada por elas. Mais importante, a análise de regressão nos forneceu uma base para fazermos *predições* a respeito do valor da variável Y baseados no conhecimento do valor correspondente da variável X, com o qual a variável Y tem um grau conhecido e significativo de correlação. Afinal de contas, sabe-se que se um conjunto de pais é mais alto do que a média, também podemos esperar que seus filhos sejam mais altos do que a média. O próprio Galton criou uma forma de quantificar as relações entre variáveis transformando os valores de cada uma delas em uma escala comum e calculando um índice ou coeficiente numérico que resumisse a força de sua relação. No entanto, foi Karl Pearson, um matemático discípulo de Galton, quem refinou o método e desenvolveu a fórmula mais amplamente usada para o cálculo dos coeficientes de correlação.

COEFICIENTES DE CORRELAÇÃO

Como vimos, o termo *correlação* simplesmente se refere ao grau pelo qual as variáveis estão relacionadas. O grau e a direção da correlação entre variáveis é medido por meio de vários tipos de *coeficientes de correlação*, que são números que podem flutuar entre −1,00 e +1,00. O quadro Consulta Rápida 2.6 lista alguns outros fatos básicos, mas, freqüentemente, mal-compreendidos, a respeito dos coeficientes de correlação em geral.

CONSULTA RÁPIDA 2.6

Três fatos essenciais a respeito da correlação em geral

1. O grau de relação entre duas variáveis é indicado pelo número do coeficiente, enquanto que a direção da relação é indicada pelo sinal.
   Um coeficiente de correlação de −0,80, por exemplo, indica exatamente o mesmo grau de relação que um coeficiente de +0,80. Seja positiva ou negativa, uma correlação é baixa na medida em que seu coeficiente se aproxima de zero. Embora estes fatos possam parecer óbvios, a natureza aparentemente convincente dos sinais negativos muitas vezes faz com que as pessoas os esqueçam.
2. A correlação, mesmo quando alta, não implica causação.
   Se duas variáveis, X e Y, estão correlacionadas, pode ser porque X causa Y, porque Y causa X ou porque uma terceira variável, Z, causa tanto X quanto Y. Este truísmo, também freqüentemente ignorado, e coeficientes de correlação de moderados a altos muitas vezes são citados como se fossem prova de uma relação causal entre as variáveis correlacionadas.
3. Correlações altas nos permitem fazer predições.
   Embora a correlação não implique causação, ela implica uma certa quantidade de variância comum ou compartilhada. O conhecimento do grau em que as coisas variam em relação umas às outras é extremamente útil. Através da análise de regressão, podemos usar dados de correlação relativos a duas ou mais variáveis para derivar equações que nos permitam prever os valores esperados de uma variável dependente (Y), dentro de uma certa margem de erro, a partir dos valores conhecidos de uma ou mais variáveis independentes (X_1, X_2...X_k), com as quais a variável dependente está correlacionada.

Ao contrário das chamadas ciências exatas, nas quais a experimentação é um modo típico de proceder, nas ciências comportamentais a capacidade de manipular variáveis muitas vezes é restrita. Por isso, as pesquisas em psicologia se valem de métodos de correlação com muita freqüência. Felizmente, a variedade de delineamentos de pesquisa e métodos de análise que podem ser aplicados aos dados cresceu imensamente com o poder e a disponibilidade dos computadores modernos. Algumas das técnicas hoje consideradas lugares-comuns para a análise simultânea de dados de múltiplas variáveis, como a regressão múltipla e a análise de trilha, são tão sofisticadas que permitem aos psicólogos e outros cientistas sociais fazer algumas inferências a respeito de relações causais com alto grau de confiança.

A técnica estatística a ser usada para computar um coeficiente de correlação depende da natureza das variáveis a serem correlacionadas, dos tipos de escalas usadas para mensurá-las e do padrão de sua relação. Mais uma vez, uma revisão completa destes métodos está além do alcance deste livro. No entanto, o índice mais amplamente usado da quantidade de correlação de duas variáveis merece alguma atenção.

Coeficiente de correlação produto-momento de Pearson

A fórmula básica criada por Karl Pearson para calcular o coeficiente de correlação de dados bivariados de uma amostra é conhecido formalmente como o *coeficiente de correlação produto-momento de Pearson*. A fórmula de definição deste coeficiente, comumente conhecida como r de Pearson, é:

$$r_{xy} = \frac{\sum xy}{N s_x s_y} \qquad (2.1)$$

onde

r_{xy} = a correlação entre X e Y;
x = o desvio de um escore X da média dos escores X;
y = o desvio de um escore Y correspondente da média dos escores Y;
S xy = a soma de todos os produtos cruzados dos desvios (isto é, a soma dos produtos de cada desvio de x vezes seu desvio de y correspondente);
N = o número de pares no conjunto de dados bivariados;
s_x = o desvio padrão dos escores X;
s_y = o desvio padrão dos escores Y.

Embora a fórmula computacional de escore bruto para o r de Pearson seja mais complicada do que a fórmula de definição, a disponibilidade de programas de computador para calcular coeficientes de correlação torna a fórmula computacional praticamente desnecessária. Por outro lado, a Fórmula (2.1) e a Fórmula (2.2), ainda mais abreviada, são um auxílio considerável na compreensão do significado

do coeficiente de correlação. O quadro Consulta Rápida 2.7 lista a notação básica para a correlação, juntamente com duas versões da fórmula do *r* de Pearson.

CONSULTA RÁPIDA 2.7

Notação básica para a correlação

Variável X	**Variável Y**
X = Um escore na variável X	Y = Um escore na variável Y
$x = X - \bar{X}$ = Escore de desvio em X	$y = Y - \bar{Y}$ = Escore de desvio em Y
S_x = Desvio padrão de X	S_y = Desvio padrão de Y
z_x = Escore padrão na variável X	z_y = Escore padrão na variável Y
$z_x = \frac{X - \bar{X}}{s_x}$	$z_y = \frac{Y - \bar{Y}}{s_y}$

Fórmulas para o *r* de Pearson:

$$r_{xy} = \frac{\sum xy}{Ns_x s_y}$$ Fórmula (2.1), fórmula de definição

$$rx_y \quad \frac{\sum Z_x Z_y}{N}$$ Fórmula (2.2), fórmula de escore-padrão

onde N = número de observações pareadas de X e Y usadas para calcular r.

Coeficiente de determinação = r^2_{xy}

O *r* de Pearson é na verdade a média dos produtos cruzados dos escores padrões das duas variáveis correlacionadas. A fórmula que corporifica esta definição é

$$r_{xy} = \frac{\sum z_x z_y}{N} \qquad (2.2)$$

onde

rx_y = a correlação entre X e Y;
Z_x = os escores padrões da variável X, obtidos dividindo-se cada escore de desvio em X pelo desvio padrão de X;
Z_y = os escores padrões da variável Y, obtidos dividindo-se cada escore de desvio em Y pelo desvio padrão de Y.

Somar os produtos cruzados dos escores z das variáveis X e Y e dividir o resultado pelo número de pares em um conjunto de dados produz uma média que reflete a quantidade de relação entre X e Y, ou seja, o *r* de Pearson.

A Fórmula (2.2) é de interesse no contexto da testagem psicológica, não apenas devido à sua brevidade e base conceitual, mas também porque serve para introduzir a noção de *escores padrões* ou *escores z*, dos quais vamos tratar outra vez no próximo capítulo. O leitor pode ter notado que os valores ao longo da linha de base da curva normal e na Tabela de Áreas da Curva Normal apresentada no Apêndice C são dados em termos de escores *z*. O motivo para isto é que um escore *z* representa a distância entre cada valor em uma distribuição e a média desta distribuição, expresso em termos da unidade de desvio padrão para esta distribuição. A fórmula (2.2) de escore padrão para o *r* de Pearson simplesmente oferece uma forma mais compacta de expressar a relação entre duas variáveis.

Condições necessárias para o uso do r de Pearson

Embora seja de longe o coeficiente de correlação mais usado, o *r* de Pearson é apropriado apenas para dados que satisfazem certas condições. Depois que Pearson desenvolveu sua fórmula, muitos métodos diferentes foram desenvolvidos para obter coeficientes de correlação para vários tipos de dados bivariados. A derivação do coeficiente de correlação produto-momento de Pearson se assenta nas seguintes premissas:

1. Os pares de observações são independentes entre si.
2. As variáveis a serem correlacionadas são contínuas e medidas em escalas de intervalo ou razão.
3. A relação entre as variáveis é linear, isto é, aproxima-se de um padrão de linha reta, como descrito anteriormente.

Se a primeira e a segunda dessas premissas ou condições são satisfeitas pode ser facilmente determinado a partir do conhecimento sobre a maneira como os dados foram coletados e o tipo de dados em questão. Se os pares de escores ou observações a serem correlacionados forem obtidos independentemente uns dos outros, a primeira premissa foi satisfeita. Se os dados para ambas as variáveis representam quantidades contínuas, a segunda premissa foi satisfeita.

Satisfazer a terceira e mais crítica premissa do *r* de Pearson requer a inspeção do gráfico de dispersão dos dados bivariados para verificar se a distribuição dos casos se enquadra na forma elíptica indicativa de uma relação linear, representada na Figura 2.6, Painel A. Quando esta premissa é violada, o *r* de Pearson não é um índice preciso de correlação.

Desvios da linearidade

Para os fins de determinar a aplicabilidade do r de Pearson a um conjunto de dados bivariados, existem dois modos como um gráfico de dispersão pode se desviar da forma elíptica que indica uma relação positiva linear. O primeiro e mais óbvio é se

existe uma inclinação significativa da forma elíptica, como na Figura 2.6, Painéis B e C. Estes desvios indicam que não há mais uma relação linear e, portanto, a relação entre X e Y não é a mesma ao longo de toda a gama de seus valores. A segunda forma como os gráficos de dispersão podem se desviar da elipse que indica uma

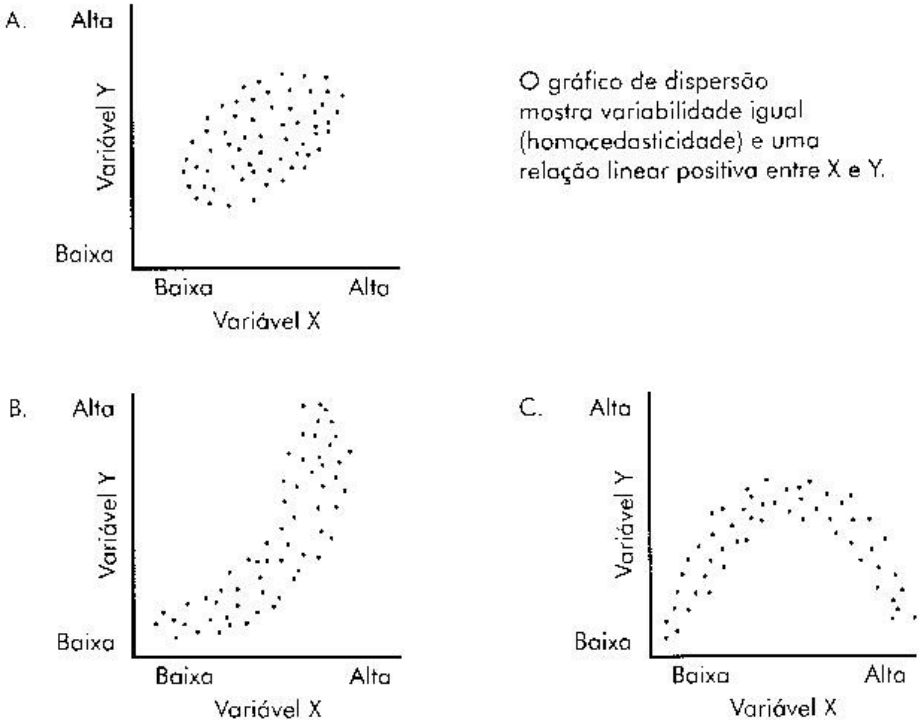

Figura 2.6 Gráficos de dispersão ilustrando várias características de dados bivariados.
Nota. Cada ponto marca a localização de um par de observações ou escores em X e Y.

relação linear é uma condição denominada *heterocedasticidade*. Isto simplesmente significa que a dispersão ou variabilidade do gráfico de dispersão não é uniforme ao longo de toda a gama de valores das duas variáveis. Para que se possa usar o coeficiente de correlação *r* de Pearson, o gráfico de dispersão precisa mostrar uma quantidade bastante uniforme de dispersão, ou *homocedasticidade*, ao longo de toda a amplitude. O gráfico de dispersão da Figura 2.6, Painel A, é homocedástico, enquanto que os Painéis D e E são heterocedásticos.

Uma das maneiras de evitar a aplicação imprópria do *r* de Pearson é produzir um gráfico de dispersão dos dados bivariados e examinar seu formato em busca de possíveis desvios da linearidade. Se a fórmula do *r* de Pearson for aplicada a dados que se desviam de uma relação linear reta, seja em termos de inclinação na forma do gráfico de dispersão ou devido à heterocedasticidade, o coeficiente de correlação resultante vai ser um índice incorreto da relação entre X e Y.

Restrição de amplitude e correlação

Uma característica importante e muitas vezes negligenciada do *r* de Pearson diz respeito ao modo como ele é afetado pela variabilidade das variáveis correlacionadas. Em termos mais simples, o efeito de uma restrição na amplitude de qualquer uma das variáveis é a redução do tamanho de *r*.

Exemplo 1: Um caso extremo. O caso mais extremo, embora não muito realista, de restrição da amplitude seria uma situação na qual não existe qualquer variabilidade em uma das variáveis correlacionadas. Se consultarmos a fórmula de definição do *r* de Pearson apresentada no quadro Consulta Rápida 2.7, podemos facilmente ver que se não houver variabilidade nos escores de X ou Y (isto é, se todos os valores de X ou Y forem os mesmos), todos os escores de desvio da respectiva variável e o numerador da fórmula do coeficiente *r* de Pearson serão zero, resultando assim em um coeficiente de correlação zero. Este é apenas um exemplo da importância singular da variabilidade enfatizada no início deste capítulo.

Exemplo 2: O efeito da restrição de amplitude na testagem de seleção para emprego. Se todas as pessoas que se candidatassem a um grande número de vagas disponíveis em uma nova empresa fossem contratadas, independentemente de seus escores em um teste de aptidão vocacional, haveria uma grande chance de encontrarmos uma correlação bastante alta entre seus escores e medidas de produtividade no trabalho, obtidas alguns meses depois de elas serem contratadas. Como podemos pressupor que um grande número de candidatos iria exibir variações bastante amplas tanto nos escores de testes de aptidão quanto em produtividade no trabalho, a relação entre a aptidão e a produtividade certamente estaria refletida no coeficiente de correlação. Se, depois de algum tempo, o processo de seleção de pessoal se tornasse mais restritivo – de tal modo que apenas aqueles candidatos que obtivessem escores altos no teste de aptidão fossem contratados – o efeito de rede desta mudança seria diminuir a amplitude de capacidades entre os funcionários recém-contratados. Por isso, se um novo coeficiente fosse calculado apenas com os dados dos recém-contratados, o grau de correlação entre os escores no teste de aptidão e a produtividade seriam reduzidos. O diagrama da Figura 2.7 representa a alta cor-

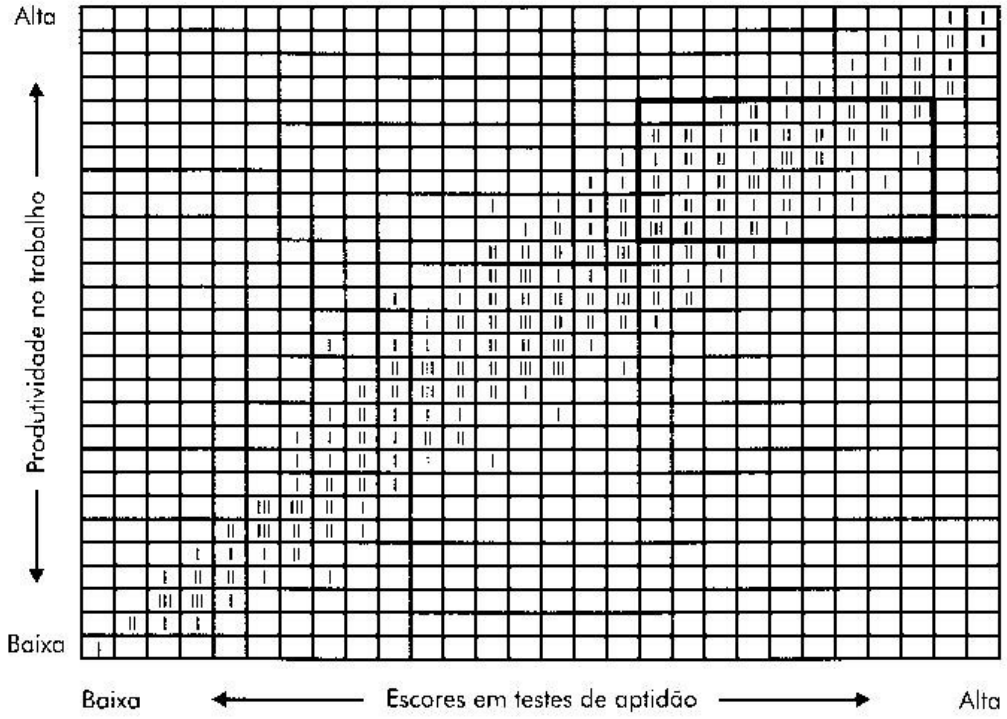

Figura 2.7 Efeito da restrição de amplitude na correlação.

relação positiva entre os escores do teste de aptidão e a produtividade no trabalho que poderia haver entre o grande grupo heterogêneo de pessoas inicialmente contratadas. O pequeno segmento na porção superior direita do diagrama representa a correlação baixa, quase inexistente, que provavelmente se constataria no grupo muito mais restrito dos candidatos mais bem colocados posteriormente contratados.

Assim como a restrição na amplitude de variáveis correlacionadas vai diminuir a correlação entre elas, uma ampla gama de variabilidade nas variáveis correlacionadas vai tender a aumentar o tamanho do coeficiente de correlação obtido e possivelmente superestimar a relação entre as duas variáveis. O fato de que os coeficientes de correlação dependem da variabilidade das amostras dentro das quais são encontrados enfatiza a importância de examinarmos sua composição do ponto de vista de seu ajustamento. Embora algumas correções estatísticas para restrições de amplitude possam ser usadas quando a variabilidade de uma amostra é reduzida, não há substituto para o cuidado em garantir que a variabilidade das amostras dentro das quais os coeficientes são calculados corresponda à variabilidade do grupo ou grupos aos quais as correlações obtidas serão aplicadas.

Outros métodos correlacionais

O *r* de Pearson pode ser usado em uma ampla gama de casos, desde que as condições necessárias sejam satisfeitas. Quando isso não acontece, outros procedimen-

tos podem ser aplicados para obter correlações para dados bivariados. Por exemplo, quando as variáveis a serem correlacionadas estão em forma ordinal, o método de correlação de escolha – já mencionado em relação às escalas ordinais – é o coeficiente de correlação para diferenças de posto de Spearman, geralmente conhecido como *rho* de Spearman *(rs)*. Se a relação entre duas variáveis for curvilinear, a razão de correlação – comumente conhecida como eta (η) – pode ser usada. Quando uma das variáveis a ser correlacionada é dicotômica, a correlação *bi-serial pontual, ou* rp_b, é usada, enquanto que se ambas as variáveis são dicotômicas, o coeficiente fi (ϕ) é empregado. As variáveis dicotômicas muitas vezes surgem na análise de itens de testes registrados em termos de aprovação-reprovação ou respostas de verdadeiro-falso.

Existem muitos outros tipos de coeficientes de correlação adequados para tipos específicos de dados, que podem ser encontrados em manuais de estatística conforme a necessidade. Uma variante particularmente importante é o *coeficiente de correlação múltipla (R)*, usado quando uma única variável dependente (Y) se correlaciona com dois ou mais preditores combinados (X_1, X_2...X_k).

CONCLUSÃO

Este capítulo apresentou os conceitos estatísticos básicos necessários para compreender escores de testes e seu significado. A estatística existe para nos auxiliar a dar sentido aos dados, mas não responde a perguntas. Para isso, precisamos usar nosso julgamento aliado à estatística. Vamos nos deparar novamente com estes conceitos no contexto dos vários aspectos técnicos dos testes – como informações normativas, fidedignidade e validade – que nos permitem avaliar sua qualidade como instrumentos de mensuração psicológica.

Teste a si mesmo

1. Quais das seguintes escalas de mensuração é a única que tem um ponto-zero significativo?
   (a) nominal
   (b) racional
   (c) ordinal
   (d) intervalar

2. Tom e Jerry pontuaram no 60° e 65° percentis, respectivamente, em um teste de habilidade de linguagem. Mary e Martha pontuaram no 90° e 95° percentis, respectivamente, no mesmo teste. Podemos concluir que a diferença entre Tom e Jerry em termos de suas habilidades de linguagem é a mesma que a diferença entre Mary e Martha. Esta afirmação é
   (a) verdadeira
   (b) falsa

3. Em um teste de capacidade cognitiva geral, uma criança de 5 anos obtém um escore de idade mental de 4 anos, e uma criança de 10 anos obtém um escore de idade mental de 9 anos. Se calculássemos seus QIs segundo a fórmula original do QI-razão, o resultado seria o seguinte:

(a) ambas as crianças obteriam o mesmo QI-razão
(b) a criança de 5 anos obteria um QI-razão mais alto
(c) a criança de 10 anos obteria um QI-razão mais alto

4. No distribuição 2, 2, 2, 2, 3, 3, 3, 8, 11, a média, a mediana e a moda são, respectivamente
   (a) 4, 3 e 2
   (b) 2, 4, e 3
   (c) 3, 4 e 2
   (d) 2, 3 e 4

5. Para a testagem e para muitos outros fins, a quintessência do índice de variabilidade em uma distribuição de escores é
   (a) a soma dos quadrados dos escores de desvio
   (b) a raiz quadrada da variância
   (c) a distância semi-interquartílica

6. Quais das seguintes afirmações a respeito do modelo da curva normal *não* é verdadeira?
   (a) ela é bilateralmente simétrica
   (b) seus limites se estendem ao infinito
   (c) sua média, mediana e moda coincidem
   (d) ela é multimodal

7. A área de uma distribuição normal entre $+1\sigma$ e -1σ engloba aproximadamente ___ da curva.
   (a) 50%
   (b) 68%
   (c) 95%
   (d) 99%

8. Se a forma da distribuição dos escores obtidos em um teste for significativamente assimétrica, isto significa que o teste é provavelmente ___ para os testandos em questão.
   (a) muito fácil
   (b) muito difícil
   (c) difícil demais ou fácil demais
   (d) correto

9. Quais dos seguintes coeficientes representa o grau mais forte de correlação entre duas variáveis?
   (a) – 0,80
   (b) – 0,20
   (c) + 0,20
   (d) + 0,60

10. Se a amplitude de valores de uma de duas variáveis que são correlacionadas usando o coeficiente de correlação produto-momento de Pearson (*r* de Pearson) é restrita, o tamanho do coeficiente obtido
    (a) será reduzido
    (b) será aumentado
    (c) não será afetado

Respostas: 1.b; 2.b; 3.c; 4.a; 5.b; 6.d; 7.b; 8.c; 9.a; 10.a.