Mais Comentários sobre Correlação

1 Causa e Efeito

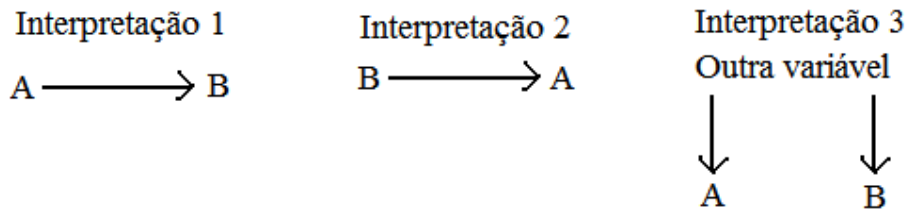
Considere as seguintes afirmações:

- Pesquisas estabeleceram que existe uma forte correlação entre o uso de punição física pelos pais e o desenvolvimento de comportamentos agressivos em seus filhos. Os pais não devem usar essa forma de impor disciplina se não quiserem que seus filhos tornem-se agressivos.
- Existe uma correlação significante entre o desmame precoce e o aparecimento de irritabilidade em crianças pequenas. Portanto, mães não devem se apressar em parar de amamentar seus bebês se elas quiserem ter filhos tranqüilos.
- Pobreza está correlacionada com crime. Portanto, famílias que têm uma renda alta têm menor probabilidade de que seus filhos sejam criminosos.

Em cada um dos casos acima, assumiu-se que a primeira variável é a *causa* da segunda. Em geral, quando há uma correlação significante entre duas variáveis, A e B, pode haver várias possíveis interpretações para a relação entre elas:

1. A variável A tem um efeito causal sobre a variável B;
2. A variável B tem um efeito causal sobre a variável A;
3. Tanto A como B estão relacionadas a alguma outra variável;
4. Apesar de significante a um nível α (por exemplo, 0,05), a correlação entre A e B não é real e o valor do coeficiente de correlação é apenas fruto de uma coincidência.

As primeiras três interpretações estão ilustradas abaixo.



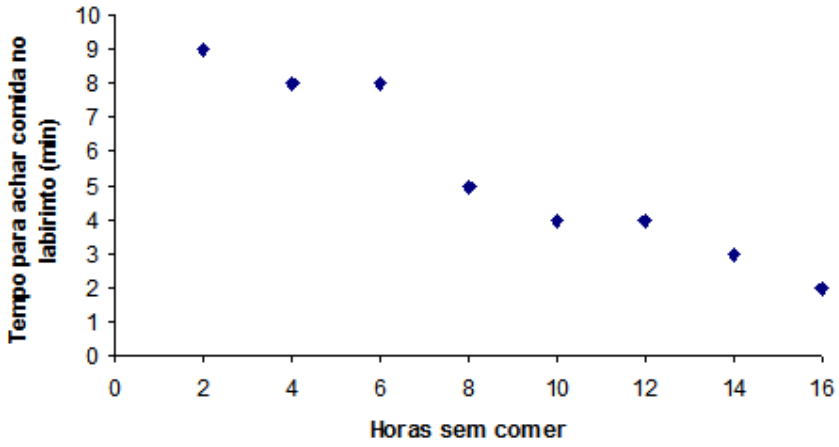
Um exemplo em que a terceira interpretação seria a correta é o da medição da temperatura em uma sala com dois termômetros, um que mede a temperatura em graus Celsius (^{o}C) e outro que mede a temperatura em graus Fahrenheit (^{o}F). As duas medições estão correlacionadas, mas não se pode dizer que um termômetro influencia o outro (eles são independentes). O que causa a correlação entre as duas medidas é a temperatura da sala.

Da mesma forma, pode-se argumentar que não é a punição física aplicada pelos pais que causa a agressividade de seus filhos, mas que as duas são conseqüência de um mesmo ambiente. Um ambiente social violento pode produzir tanto pais que batem em seus filhos como filhos agressivos, e não seriam as punições físicas dos pais que tornariam seus filhos agressivos.

Em geral, é uma boa prática questionar se, ao invés de ser A que causa B, não seria B que causa A. Talvez não seja a punição física que cause filhos agressivos, mas filhos com agressividade inata que levem seus pais a usarem de violência contra eles. Da mesma forma, talvez sejam os bebês que nasçam irritadiços que induzam suas mães a abandonar o aleitamento materno.

Em alguns casos, a configuração do problema permite que se determine se uma variável causa a outra. Um exemplo disso é quando uma variável *antecede* a outra. Por exemplo, se de fato existir uma correlação positiva entre a altura da pessoa e o seu sucesso profissional, a altura seria a causa do sucesso e não o contrário. Pode até ser que o sucesso profissional de uma pessoa a faça parecer mais alta segundo a *percepção* de outros, por exemplo, pesquisas nos Estados Unidos indicam que a população tende a superestimar a altura do candidato que ganha uma disputa para a presidência. Mas o sucesso profissional não tem como influenciar a altura de uma pessoa, pois esta é uma característica determinada pela genética e pela alimentação da pessoa na infância, antes de ela ingressar no mercado de trabalho. A interpretação 3, no entanto, poderia ser válida neste caso. Outras características genéticas e do ambiente onde a pessoa cresce poderiam causar tanto uma alta estatura como um sucesso profissional.

Uma outra situação onde se pode determinar qual variável é a causa da outra é quando se faz um *experimento controlado* em laboratório. Por exemplo, quando se mede o efeito que a quantidade de horas sem alimentação tem sobre o tempo que um camundongo leva para achar onde está a comida em um labirinto, pode-se obter um diagrama de dispersão como o abaixo.



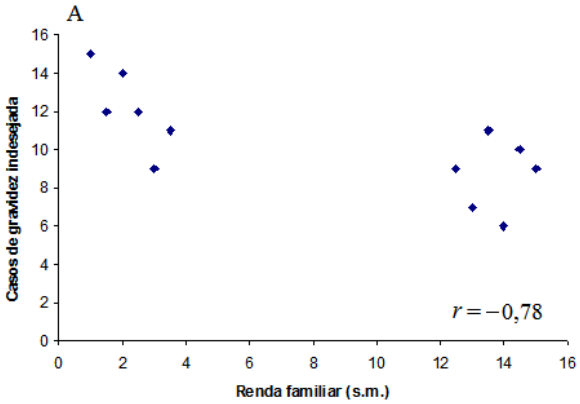
Neste caso, é a quantidade de horas sem comer que causa uma diminuição no tempo gasto pelo animal para achar comida no labirinto. Em situações experimentais como a mostrada acima, costuma-se chamar a variável causadora (colocada no eixo-*x*) de variável *independente* e a outra variável (colocada no eixo-*y*) de variável *dependente*.

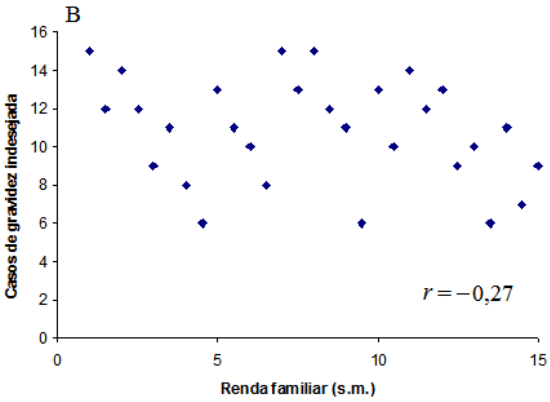
Um outro exemplo é o de um experimento em que se mostra uma seqüência de palavras separadas por um certo intervalo de tempo a uma pessoa e depois pede-se a ela que repita as palavras mostradas. Em um experimento desse tipo, quando se aumenta o intervalo de tempo entre as palavras, o índice de acertos da pessoa em se lembrar das palavras mostradas também aumenta. Pelo *desenho* do experimento, as palavras são mostradas primeiro, com o intervalo de tempo entre elas determinado pelo experimentador, e só depois é que se pede à pessoa para se lembrar das palavras. Portanto, o intervalo de tempo é a variável independente e o índice de acertos (a fração de palavras corretamente lembradas) é a variável dependente.

2 Escolha da Faixa de Variação dos Valores

Às vezes, a escolha dos dados para serem incluídos em um estudo de correlação pode fazer parecer que existe uma forte correlação entre eles, quando de fato ela não existe.

Por exemplo, considere um estudo feito para medir a correlação entre a faixa de renda familiar e o número de casos de gravidez indesejada. Suponha que o estudo acabe concluindo que existe uma forte correlação negativa entre as duas, com base no diagrama de dispersão da esquerda (indicado por A) abaixo.

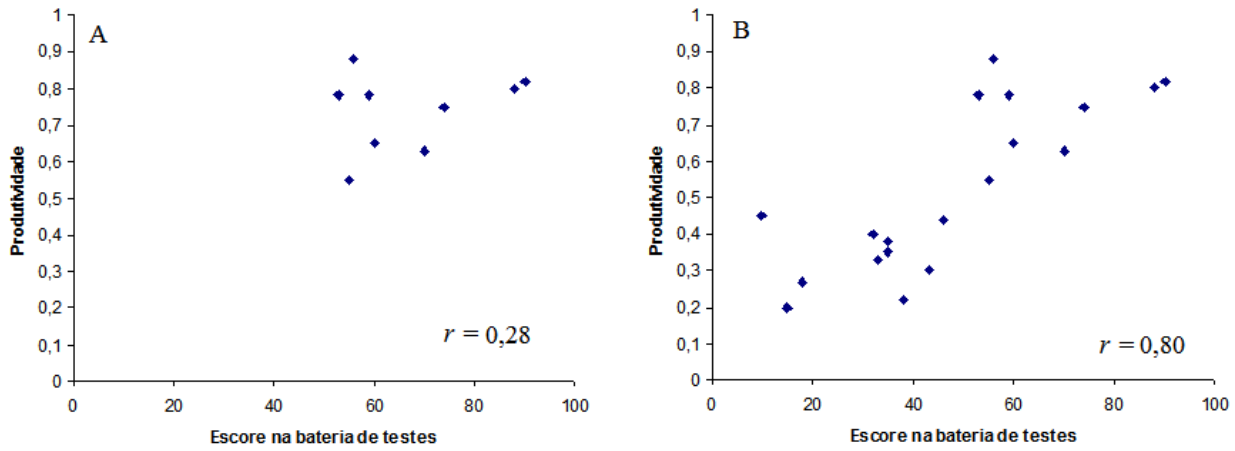




Um resultado como esse poderia ser usado politicamente para se dizer que as pessoas pobres são menos cuidadosas em suas relações sexuais e que, portanto, há uma maior incidência de casos de gravidez indesejada entre elas. Dependendo da sociedade em que isso ocorra, uma tal conclusão poderia levar a uma campanha em favor de uma melhor educação sexual entre os pobres, ou a uma campanha (explícita ou oculta) de esterilização em massa de mulheres pobres.

Observando o diagrama da figura A, porém, vemos que ele se baseia em uma amostra *enviesada*, que só leva em conta famílias com renda familiar alta e baixa, desprezando as de renda intermediária. Essa amostragem seletiva implica em uma forte correlação, mas uma amostragem mais representativa da população (veja o diagrama da figura B) poderia implicar em uma correlação mais fraca.

Um efeito oposto pode ocorrer quando a faixa de valores é escolhida de uma maneira diferente. Suponha que uma companhia contrate um psicólogo para ajudá-la a selecionar candidatos a empregos através de uma bateria de testes psicológicos. Suponha que, um ano após a contratação do grupo de candidatos selecionados, a companhia resolva fazer um estudo de correlação entre a produtividade desses empregados e o seu escore no teste aplicado pelo psicólogo e obtenha uma baixa correlação como a mostrada no diagrama A da figura abaixo.



A companhia poderia, então, dispensar os serviços do psicólogo alegando que a bateria de testes utilizada por ele não é boa para prever se o candidato terá alta produtividade ou não. O psicólogo, porém, poderia se defender dessa crítica alegando que, caso os candidatos que tiveram escores baixos fossem contratados, suas produtividades seriam menores e teríamos um diagrama como o mostrado na figura B, indicando uma correlação alta entre resultado no teste e produtividade.

3 Relações Não-Lineares (funções de potência e exponencial)

Os coeficientes de correlação de Pearson e de Spearman são usados para medir a força da correlação linear entre duas variáveis *X* e *Y*. Em geral, eles não têm muita utilidade quando a relação entre *X* e *Y* não é linear. Porém, há dois tipos importantes de relações não-lineares que podem ser *transformados* em relações lineares com o uso da função logaritmo: as funções de potência e exponencial (veja as notas da aula especial sobre funções e seus gráficos).

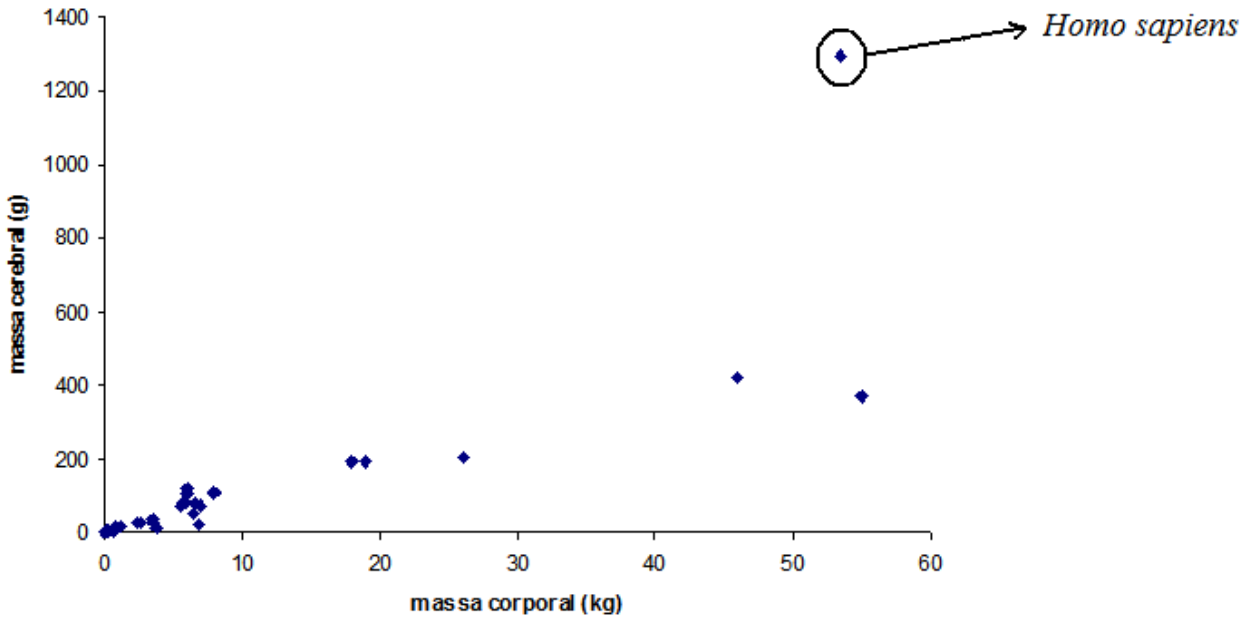
Quando estamos desconfiados que a relação entre duas variáveis é de um desses dois tipos, devemos fazer gráficos de *Y versus X* dos tipos log-log e linear-log para checar se a relação fica linear em um deles. Para fazer essa checagem, devemos calcular o coeficiente de correlação de Pearson.

Vamos ver um exemplo. A tabela abaixo mostra valores da massa corporal (kg) e da massa cerebral (g) para 37 espécies de primatas, incluindo os humanos. Os dados foram tirados do seguinte artigo de revisão: Leonard, W.R., Robertson, M.L., Snodgrass, J.J. and Kuzawa, C.W., Metabolic correlates of hominid brain evolution. *Comparative Biochemistry and Physiology Part A*, 136: 5-15, 2003.

Espécie	**massa corporal (kg)**	**massa cerebral (g)**
Arctocebus calabarensis	0,323	7,2
Cheirogaleus medius	0,177	3,1
Eulemur fulvus	2,397	25,2
Euoticus elegantulus	0,274	7,2
Galago senegalensis	0,186	4,8
Galagoides demidoff	0,081	3,4
Lemur catta	2,678	25,6
Lepilemur ruficaudatus	0,682	7,6
Loris tardigradus	0,322	6,6
Microcebus murinus	0,054	1,8
Nycticebus coucang	0,8	12,5
Otolemur crassicaudatus	0,85	10,3
Perodicticus potto	1,15	14
Propithecus verreauxi	3,48	26,7
Varecia variegata	3,512	34,2
Alouatta palliata	6,4	51
Aotus trivirgatus	0,85	16
Callithrix geoffroyi	0,28	7,6
Callitrix jacchus	0,28	7,6

Cebuella pygmaea	0,14	4,5
Cercopitecus mitis	6,5	76
Cercocebus torquatus	7,9	104
Colobus guereza	7	73
Erythrocebus patas	6	118
Homo sapiens	53,5	1295
Hylobates lar	6	102
Macaca fascicularis	5,5	74
Macaca fuscata	5,9	84
Macaca mulatta	8	110
Pan troglodytes	46	420
Papio anubis	26	205
Papio cynacephalus	19	195
Papio papio	18	190
Papio ursinus	18	190
Pongo pymaeus	55	370
Saguinus geoffroyi	3,8	10
Saimiri sciureus	6,8	22

O gráfico (diagrama de dispersão) da massa cerebral *versus* a massa corporal para os dados da tabela está dado abaixo.

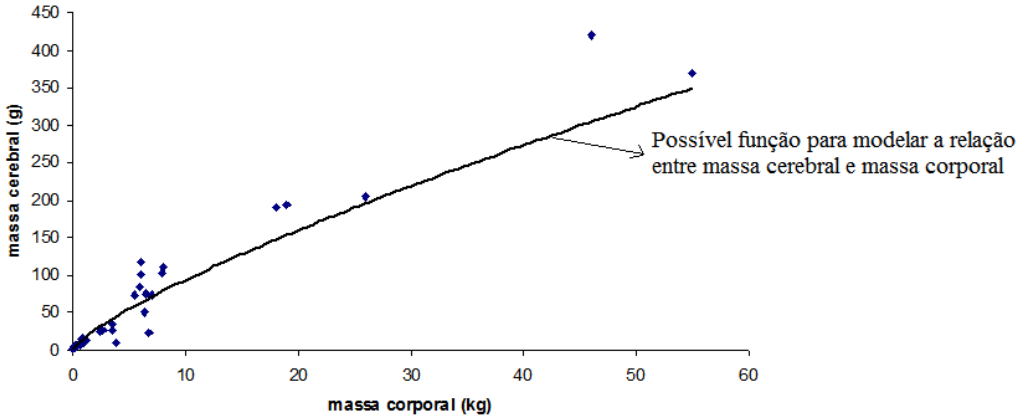


Vemos que os dados parecem indicar uma relação não-linear entre a massa cerebral e a massa corporal dos primatas, com o caso do ser humano sendo um ponto fora da curva: a massa do cérebro do ser humano é muito maior do que o valor esperado para um primata com a sua massa corporal.

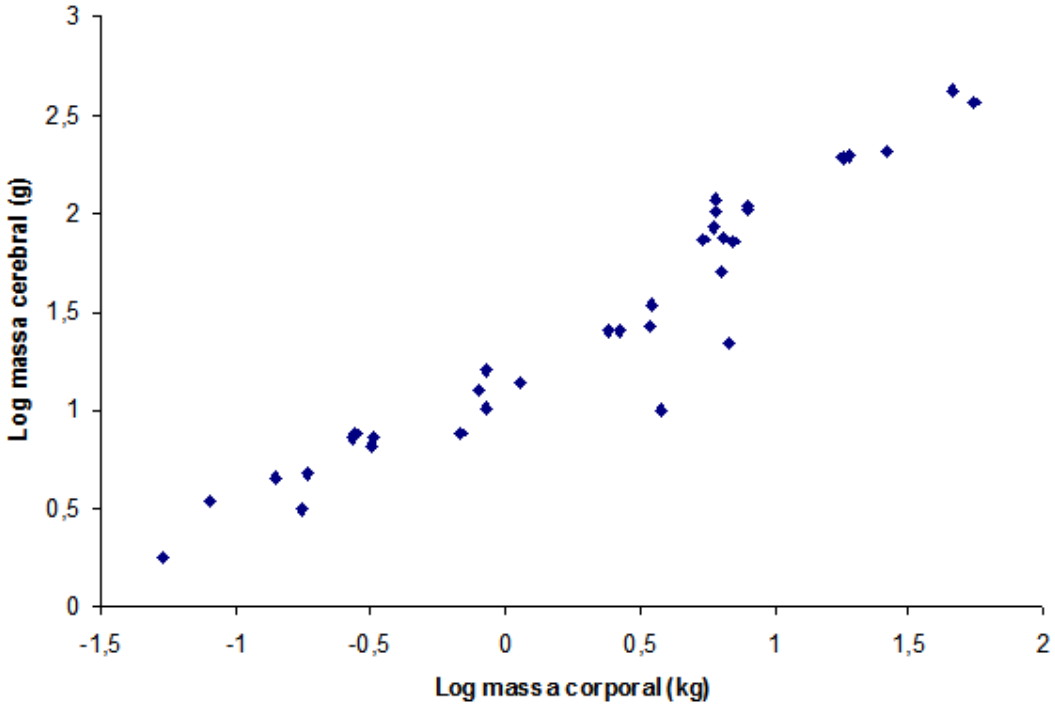
Desprezando o dado para o ser humano, uma possível função capaz de modelar a relação entre as duas variáveis é uma função de potência como a mostrada abaixo com expoente *p* menor do que 1,

$$\text{massa cerebral} = \text{k.}(\text{massa corporal})^p \,.$$

O gráfico dessa função de potência com $p < 1$ seria algo como o mostrado no desenho abaixo.



Para verificar se uma função de potência realmente serve para modelar a relação entre massa cerebral e massa corporal, devemos tomar os logaritmos das duas variáveis e fazer um gráfico de log(massa cerebral) contra log(massa corporal). O resultado está mostrado no gráfico abaixo.



Parece que os pontos no gráfico acima são muito bem ajustados por uma linha reta. Para nos certificarmos de que uma linha reta realmente fornece um bom modelo para a relação entre o log da massa cerebral e o log da massa corporal, devemos calcular o coeficiente de correlação de Pearson.

Fazendo o cálculo (por exemplo, usando a função "correl" do Excel), obtemos que o valor de r é 0,96. Esta é uma correlação linear bem forte. Fazendo o teste de significância para este resultado (veja a aula 3), temos que gl = 36 – 2 = 34 (lembre-se que estamos desprezando a espécie humana) e que $t_0 = r\sqrt{(n-2)/(1-r^2)} = 20{,}09.$ Este valor de t_0 é muito maior que t(gl), não só para α = 0,05 como também para α = 0,005. Portanto, podemos concluir que ele é significante.

Isto indica que o modelo linear fornece um bom ajuste para a relação entre os logaritmos da massa cerebral e da massa corporal. Portanto, a relação entre as variáveis originais (isto é, sem os logaritmos) é muito bem aproximada por uma função de potência.

Uma área da psicologia em que funções de potência e exponenciais são muito usadas para modelar dados experimentais é a que estuda o aprendizado. Em particular, as chamadas "curvas de aprendizado", em que se relaciona o tempo gasto para executar uma dada tarefa e o número de vezes em que ela foi repetida (a prática da pessoa na execução da tarefa) costumam ser ajustadas por funções de potência ou exponenciais. A este respeito vejam, por exemplo, o artigo anexo: Ritter, F. E. and Schooler, L. J., The learning Curve. In: *International Encyclopedia of the Social & Behavioral Sciences*.

4 Usos comuns de coeficientes de correlação em psicologia

As situações mais comuns em que estudos de correlação são feitos em psicologia são as seguintes:

- Estudos não experimentais: De longe, o uso mais comum de coeficientes de correlação em psicologia ocorre em estudos em que duas variáveis *já existentes* são medidas para uma amostra. Tais estudos são chamados em psicologia de "não-experimentais" (ou de "correlacionais" por alguns autores), para diferenciar de estudos experimentais controlados em que uma variável independente é manipulada para *causar ou não* variações em uma variável dependente. Exemplos de estudos não-experimentais de correlação são: quantidade de fumo ingerida e nível de ansiedade; atitudes sexistas e racistas de pessoas; horas que uma criança passa assistindo a programas violentos na TV e nível de agressividade.

- Testes de confiabilidade: Testes desse tipo são aplicados para, por exemplo, determinar se alguma medida feita em uma amostra de pessoas é confiável para ser usada ao longo do tempo. O método usado neste caso é o do teste e re-teste: por exemplo, toma-se uma amostra de n pessoas em um dado momento e mede-se alguma variável para elas; um tempo depois, digamos, seis meses, mede-se a mesma variável para o mesmo grupo de pessoas e faz-se um estudo de correlação entre os dois conjuntos de medidas. Testes desse tipo são importantes, por exemplo, para se avaliar a confiabilidade dos julgamentos de pessoas responsáveis por atribuir escores ou notas a outras.
- Estudos com gêmeos: Gêmeos idênticos (monozigóticos) ou, eventualmente, fraternos (dizigóticos) formam *pares ideais* para estudos de correlação. De fato, é comum que medidas ou escores para gêmeos desses tipos sejam correlacionados. Exemplos são os estudos sobre influências hereditárias no comportamento, dos quais os mais famosos são os estudos sobre a correlação entre os QIs de gêmeos idênticos e fraternos criados separadamente, isto é, em ambientes diferentes. Para mais detalhes sobre estudos de correlação entre desempenhos intelectuais ou traços de personalidade de gêmeos, ver, por exemplo, o Capítulo 3 do livro de David R. Shaffer, *Psicologia do Desenvolvimento: Infância e Adolescência*, Editora Pioneira Thomson Learning, São Paulo, 2005.