☑ 8
ANALISANDO A ASSOCIAÇÃO ENTRE AS VARIÁVEIS CATEGÓRICAS

Lembre que dissemos que existe uma **associação** entre duas variáveis se a distribuição da variável resposta muda de alguma forma à medida que a variável explicativa muda. Na comparação de dois grupos, existe uma associação se as médias da população ou proporções da população diferirem entre os grupos.

Este capítulo apresenta métodos para detectar e descrever associações entre duas variáveis categóricas. Os métodos deste capítulo nos ajudam a responder a uma pergunta como: "Existe alguma associação entre felicidade e religiosidade?". Os métodos do Capítulo 7 para comparar duas proporções são casos especiais daqueles considerados aqui nos quais ambas as variáveis têm somente duas categorias.

A Seção 8.1 introduz a terminologia para análise de dados categóricos e define a *independência estatística*, um tipo de falta de associação. A Seção 8.2 apresenta um teste de significância para determinar se duas variáveis categóricas estão associadas, e a Seção 8.3 segue com esse teste com uma *análise dos resíduos* que descreve a natureza da associação. A Seção 8.4 mostra como determinar se a associação é forte o suficiente para ter importância prática. As Seções 8.5 e 8.6 apresentam análises especializadas para variáveis ordinais.

8.1 TABELAS DE CONTINGÊNCIA

Os dados para a análise de variáveis categóricas são apresentados em **tabelas de contingência**. Este tipo de tabela exibe o número de sujeitos observados em todas as combinações de possíveis resultados para as duas variáveis.

EXEMPLO 8.1 Diferenças nas crenças políticas por gênero

Nos últimos anos nos Estados Unidos comentaristas políticos têm discutido se existe uma "diferença por gênero" nas crenças políticas. As mulheres e homens tendem a diferir sobre o seu pensamento político e comportamento eleitoral? Para investigar isso, estudamos a Tabela 8.1, da PSG de 2004. As variáveis categóricas são gênero e identificação partidária ("SEX" e "PARTYID" na PSG). As pessoas indicaram se elas se identificam mais fortemente com o partido Democrata, Republicano ou com o Independente.

A Tabela 8.1 contém respostas para 2771 entrevistados, com classificação cruzada pelo seu gênero e identificação partidária. A Tabela 8.1 é chamada de tabela de contingência 2 × 3 (lê-se "2 por 3"), significando que ela tem duas linhas e três colunas. Os totais das linhas e os totais das colunas são chamados de **distribuições marginais**. A distribuição marginal amostral para a identificação partidária, por exemplo, é o conjunto de frequências marginais (959, 991, 821). ■

☑ **Tabela 8.1** Identificação partidária (ID) e gênero, para os dados da PSG

	Identificação partidária			
Gênero	Democrata	Independente	Republicano	Total
Mulheres	573	516	422	1511
Homens	386	475	399	1260
Total	959	991	821	2771

Comparações de percentuais

Construir uma tabela de contingência de um arquivo de dados é o primeiro passo para a investigação de uma associação entre duas variáveis categóricas. Para estudar como a identificação partidária depende do gênero, convertemos as frequências a percentuais dentro de cada linha, como mostra a Tabela 8.2. Por exemplo, a proporção de 573/1511 = 0,38 ou 38% das mulheres identificam-se como Democratas. O percentual de homens que se identificam como Democratas é igual a 31% (386 de 1260). Parece que é mais provável que as mulheres se identifiquem como Democratas do que os homens.

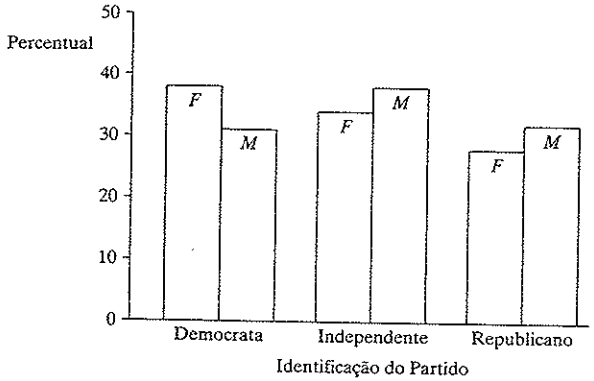
Os dois conjuntos de percentuais para mulheres e homens são chamados de **distribuições condicionais** na identificação partidária. Elas se referem à distribuição dos dados amostrais sobre identificação do partido, *condicional* ao gênero. A distribuição condicional das mulheres sobre a identificação partidária é o conjunto dos percentuais (38, 34, 28) para (Democrata, Independente, Republicano). Os percentuais somam 100 em cada linha, a menos do arredondamento. A Figura 8.1 exibe graficamente as duas distribuições condicionais.

De uma forma similar, poderíamos calcular as distribuições condicionais de gênero para cada identificação partidária. A primeira coluna iria indicar que 60% dos Democratas são mulheres e 40% são homens. Na prática, é padrão formar a distribuição condicional para a variável resposta, dentro das categorias da variável explicativa. Neste exemplo, a identificação partidária é a variável resposta, assim a Tabela 8.2 relata os percentuais dentro das linhas, que nos informa o percentual de (Democratas, Independentes, Republicanos) para cada gênero.

Outra forma de apresentar os percentuais fornece um único conjunto para todas as células da tabela usando todo o tamanho da amostra como a base. Para ilustrar, na Tabela 8.1, dos 2771 sujeitos, 573 ou 21% estão na célula (Mulher, Democrata), 386 ou 14% estão na célula (Homem, Democrata) e assim por diante. Esta distribuição do percentual é chamada de **distribuição conjunta** da amostra. Ela é útil para a comparação das frequências relativas de ocorrências para combinações dos níveis da variável. Quando fazemos a distinção entre variáveis

☑ **Tabela 8.2** Identificação partidária e gênero: percentuais calculados dentro das linhas da Tabela 8.1

	Identificação partidária				
Gênero	Democrata	Independente	Republicano	Total	*n*
Mulheres	38%	34%	28%	100%	1511
Homens	31%	38%	32%	101%	1260



☑ **Figura 8.1** Representação das distribuições condicionais da identificação partidária, na Tabela 8.2, para mulheres e homens.

respostas e explicativas, entretanto, as distribuições condicionais são mais informativas do que as distribuições conjuntas.

Diretrizes para formar tabelas de contingência

Aqui estão algumas diretrizes para quando encontrar proporções ou percentuais em tabelas de contingência. Primeiro, como já foi mencionado, encontre-as para a variável resposta dentro das categorias da variável explicativa. Construiremos a tabela de forma que a coluna da variável seja a variável resposta, como na Tabela 8.1. Assim, encontramos as proporções dentro de cada linha, dividindo a frequência de cada célula pelo total da linha.

Segundo, identifique claramente as variáveis e suas categorias e dê um título à tabela que identifica as variáveis e outras informações importantes. Terceiro, inclua o tamanho total da amostra na qual os percentuais e proporções estão baseadas. Desta forma, os leitores podem determinar as frequências das células, se elas não estiverem listadas, e eles podem encontrar os erros padrão para analisar a precisão das estimativas da proporção amostral.

Independência e dependência

Se uma associação existe na Tabela 8.1 é uma questão de se mulheres ou homens diferem nas suas distribuições condicionais quanto à identificação partidária. Respondemos à pergunta "A identificação partidária está associada ao gênero?" com referência aos conceitos de *independência* e *dependência* estatística.

☑ **Independência e dependência estatística**

Duas variáveis categóricas são **estatisticamente independentes** se as distribuições condicionais da população em uma delas são idênticas a cada categoria da outra. As variáveis são **estatisticamente dependentes** se as distribuições condicionais não são idênticas.

Em outras palavras, duas variáveis são estatisticamente independentes se o percentual da população em qualquer categoria em particular de uma variável é o mesmo para todas as categorias da outra variável. Na Tabela 8.2, as duas distribuições condicionais não são idênticas. Mas aquela tabela descreve uma amostra e a definição de independência estatística se

refere à população. Se essas observações fossem toda a população, então as variáveis seriam estatisticamente dependentes.

Para simplificar, geralmente usamos o termo *independente* em vez de *estatisticamente independente*. A Tabela 8.3 é uma tabela de contingência hipotética mostrando independência. A tabela contém os dados da população para duas variáveis – identificação partidária e grupo étnico.

O percentual de Democratas é o mesmo para cada grupo étnico, 44%. Da mesma forma, o percentual de Independentes e o percentual de Republicanos são os mesmos para cada grupo étnico. A probabilidade de que uma pessoa tenha uma identificação partidária particular é a mesma para cada grupo étnico, portanto, a identificação partidária é independente do grupo étnico.

A independência estatística é uma propriedade simétrica entre duas variáveis: se as distribuições condicionais dentro das linhas são idênticas, então assim também serão as distribuições condicionais dentro das colunas. Na Tabela 8.3, por exemplo, você pode verificar que a distribuição condicional dentro de cada coluna é igual a (74%, 7%, 19%).

EXEMPLO 8.2 O que está associado à crença na vida após a morte?

Em Pesquisas Sociais Gerais recentes, o percentual de norte-americanos que expressam a crença na vida após a morte (variável "AFTERLIF" na PSG) tem sido de aproximadamente 80%. Isso tem sido verdadeiro tanto para mulheres quanto para homens e verdadeiro para aqueles que classificam a sua raça como negra, branca ou outra. Desta forma, parece que a crença na vida após a morte pode ser estatisticamente independente das variáveis como gênero e raça. Por outro lado, enquanto aproximadamente 80% dos católicos e protestantes acreditam na vida após a morte, somente 40% dos judeus e 50% daqueles sem religião acreditam na vida após a morte. Não podemos ter certeza, não tendo dados de toda a população, mas parece que a crença na vida após a morte e religião são estatisticamente dependentes. ■

8.2 TESTE DE INDEPENDÊNCIA QUI-QUADRADO

A Tabela 8.1 contém dados amostrais. A definição de independência estatística se refere à população. Duas variáveis são independentes se as distribuições condicionais da *população* na variável resposta são idênticas. Visto que a Tabela 8.1 se refere a uma amostra, ela fornece evidência, mas não responde definitivamente se a identificação partidária e o gênero são independentes. Mesmo se eles são independentes, não esperaríamos que as distribuições condicionais da *amostra* fossem independentes. Devido à variabilidade amostral, esperamos que os percentuais da amostra difiram dos percentuais da população.

A seguir verificaremos se é plausível que a identificação partidária e o gênero sejam independentes. Se eles são realmente independentes, poderíamos esperar diferenças amostrais como as que a Tabela 8.2 mostra entre mulheres e homens nas suas distri-

☑ **Tabela 8.3** Classificação cruzada da população exibindo independência estatística. A distribuição condicional é a mesma em cada linha (44%, 14%, 42%)

	Identificação partidária			
Grupo Étnico	Democrata	Independente	Republicano	Total
Branco	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Negro	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Hispânico	110 (44%)	35 (14%)	105 (42%)	250 (100%)

buições amostrais meramente pela variação amostral? Ou diferenças deste tamanho são improváveis? Para tratar disto com um teste de significância, testamos o seguinte:

H_0: as variáveis são independentes.
H_a: as variáveis são dependentes.

O teste requer aleatorização – por exemplo, amostragem aleatória ou um experimento aleatorizado. O tamanho da amostra deve ser grande, satisfazendo a condição anunciada anteriormente nesta seção.

Frequências esperadas para a independência

O teste qui-quadrado compara frequências observadas na tabela de contingência com valores que satisfaçam a hipótese nula de independência. A Tabela 8.4 mostra as frequências observadas da Tabela 8.1 com os valores (em parênteses) que satisfazem H_0. Esses valores H_0 têm os mesmos totais das linhas e das colunas que as frequências observadas, mas satisfazem a independência. Eles são chamados de **frequências esperadas**.

☑ **Frequências esperadas e observadas**

Considere f_0 a representação de uma frequência **observada** em uma célula da Tabela. Considere f_e uma frequência **esperada**. Esta é a frequência esperada em uma célula se as variáveis fossem independentes. Ela é igual ao produto do total da linha pelo total da coluna para aquela célula, dividido pelo tamanho total da amostra.

Por exemplo, a célula no canto superior esquerdo se refere às mulheres que se identificam como Democratas. Para esta célula, $f_0 = 573$. A frequência esperada é $f_e = (1511)(959)/2771 = 522{,}9$, o produto do total da linha das mulheres pelo total da coluna dos Democratas, divido pelo tamanho total da amostra.

Vamos ver por que esta regra é lógica. Em toda a amostra, 959 de 2771 pessoas (34,6%) se identificam como Democratas. Se as variáveis fossem independentes, esperaríamos que 34,6% dos homens e 34,6% das mulheres se identificassem como Democratas. Por exemplo, 34,6% das 1511 Mulheres deveriam estar classificadas na categoria Democrata. A frequência esperada para a célula é, então:

$$f_e = \left(\frac{959}{2771}\right)1511 = 0{,}346(1511) = 522{,}9.$$

Estatística-teste qui-quadrado

A estatística-teste para H_0: independência, resume quão próximo as frequências esperadas estão das frequências observadas. Simbolizada por χ^2, ela é chamada de **estatística qui-quadrado**. Ela é igual a:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}.$$

A soma é feita em todas as células da tabela de contingência. Para cada célula, elevamos ao quadrado a diferença entre as frequências observadas e as esperadas e, então, dividimos o quadrado pela frequência esperada. Esta é a estatística-teste mais

☑ **Tabela 8.4** Identificação partidária por gênero, com as frequências esperadas em parênteses

	Identificação partidária			
Gênero	Democrata	Independente	Republicano	Total
Mulheres	573 (522,9)	516 (540,4)	422 (447,7)	1511
Homens	386 (436,1)	475 (450,6)	399 (373,3)	1260
Total	959	991	821	2771

antiga em uso hoje; ela foi introduzida pelo estatístico britânico Karl Pearson em 1900.

Quando H_0 é verdadeira, f_0 e f_e tendem a estar próximos para cada célula e o χ^2 é relativamente pequeno. Se H_0 é falsa, pelo menos alguns valores de f_0 e f_e tendem a não estar próximos, levando a valores maiores de $(f_0 - f_e)$ e a uma estatística-teste grande. Quanto maior o valor χ^2, maior a evidência contra H_0: independência.

Substituindo os valores f_0 e f_e da Tabela 8.2 na fórmula do χ^2, tem-se:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(573 - 522{,}9)^2}{522{,}9} + \frac{(516 - 540{,}4)^2}{540{,}4} + \frac{(422 - 447{,}7)^2}{447{,}7} + \frac{(386 - 436{,}1)^2}{436{,}1} + \frac{(475 - 450{,}6)^2}{450{,}6} + \frac{(399 - 373{,}3)^2}{373{,}3} = 4{,}8 + \cdots + 1{,}8 = 16{,}2.$$

O cálculo é trabalhoso, mas é simples de se obter o χ^2 utilizando um *software*. A seguir, estudaremos como interpretar sua magnitude.

A distribuição qui-quadrado

A distribuição amostral da estatística-teste χ^2 indica quão grande o χ^2 deve ser antes que exista forte evidência de que H_0 é falsa. Para tamanhos da amostra grandes, a distribuição amostral é a **distribuição de probabilidade qui-quadrado**. O nome do teste e o símbolo para a estatística-teste se referem ao nome da distribuição amostral. Aqui estão as propriedades principais da distribuição qui-quadrado:

- Ela está concentrada na parte positiva dos números reais. A estatística-teste χ^2 não pode ser negativa, visto que ela soma as diferenças ao quadrado divididas pelas frequências positivas esperadas. O valor mínimo possível, $\chi^2 = 0$, ocorreria se $f_0 = f_e$ em cada célula.
- Ela é assimétrica à direita.
- A forma precisa da distribuição depende dos **graus de liberdade** (gl). A média é $\mu = gl$ e o desvio padrão é $\sigma = \sqrt{2gl}$. Portanto, a distribuição tende a se deslocar para a direita e tornar-se mais dispersa para os gl maiores. Além disto, à medida que os gl aumentam, diminui a assimetria e a curva qui-quadrado se aproxima de uma normal. Veja a Figura 8.2.
- Para testar H_0: independência com a tabela tendo l linhas e c colunas,

$$gl = (l - 1)(c - 1).$$

Para uma tabela 2×3, $l = 2$ e $c = 3$ e $gl = (2 - 1)(3 - 1) = 1 \times 2 = 2$. Números maiores de linhas e colunas produzem valores gl maiores. Visto que tabelas maiores têm mais termos na soma para a estatística-teste χ^2, os valores χ^2 também tendem a ser maiores.

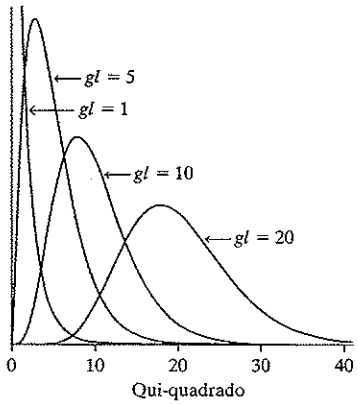
- Quanto maior o valor do χ^2, maior a evidência contra H_0: independência. O valor-p é igual à probabilidade da cauda direita acima do valor χ^2 observado. Ele mensura a probabilidade, presumindo que H_0 é verdadeira, de que χ^2 é pelo menos tão grande quanto o valor observado. A Figura 8.3 descreve o valor-p.

A Tabela C (página 652) no final do livro lista os valores do qui-quadrado para várias probabilidades da cauda direita. Estes são valores da estatística-teste χ^2 que têm valores-p iguais àquelas probabilidades. Por exemplo, a Tabela C relata que quando $gl = 2$, $\chi^2 = 5{,}99$ tem um valor-$p = 0{,}05$ e $\chi^2 = 9{,}21$ tem um valor-$p = 0{,}01$.

EXEMPLO 8.3 Estatística qui-quadrado para identificação partidária e gênero

Para aplicar o teste qui-quadrado à Tabela 8.4, testamos o seguinte:

H_0: Identificação Partidária e gênero são independentes.



☑ **Figura 8.2** A distribuição qui-quadrado. A curva tem média e desvio padrão maior à medida que os graus de liberdade aumentam.

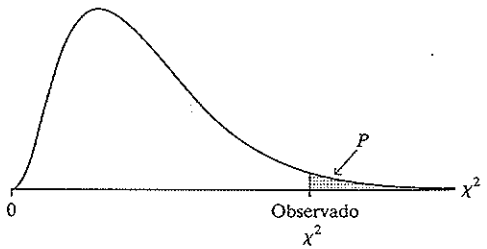
H_a: Identificação Partidária e gênero são dependentes.

Anteriormente, obtivemos a estatística-teste $\chi^2 = 16{,}2$. Na Tabela C, para $gl = 2$, 16,2 está acima de 13,82, o valor do qui-quadrado tendo uma probabilidade da cauda direita de 0,001. Portanto, concluímos que $p < 0{,}001$. O *software* indica que $p = 0{,}0003$, o que fornece uma evidência extremamente forte contra H_0. Parece provável que a identificação partidária e gênero estejam associados na população. Se as variáveis forem independentes, seria altamente incomum para uma amostra aleatória ter uma estatística χ^2 tão grande. ■

Requisitos para o tamanho da amostra

O teste qui-quadrado, como os testes z para uma e duas amostras para proporções, é um teste para amostras grandes. A distribuição do qui-quadrado é a distribuição amostral para a estatística-teste χ^2 somente se o tamanho da amostra é grande. Um procedimento simples para este requisito é que a frequência esperada f_e deva exceder 5 em cada célula. De outra forma, a distribuição do qui-quadrado pode se aproximar muito pouco da real distribuição da estatística χ^2.

Para tabelas de contingência 2×2, um teste amostral pequeno de independência é o **teste exato de Fisher**, discutido na Seção 7.5 (página 225). Esse teste se estende



☑ **Figura 8.3** O valor-p para o teste qui-quadrado de independência é uma probabilidade da cauda direita, acima do valor observado da estatística-teste.

às tabelas de tamanho arbitrário $l \times c$, mas requer um *software* especializado como o SAS (a opção EXACT em PRO FREQ) ou SPSS (o módulo Exact).

Se você tiver esse *software*, você pode usar o teste exato para *qualquer* tamanho de amostra. Você não precisa usar a aproximação qui-quadrado. Para a Tabela 8.4, o teste exato também dá o valor-p = 0,0003. A Tabela 8.5 resume as cinco etapas do teste qui-quadrado.

Usando *software* para realizar testes qui-quadrado

O teste qui-quadrado de independência é bastante trabalhoso quanto aos cálculos, portanto você terá de usar um *software* para realizá-lo. A Tabela 8.6 ilustra o resultado para a Tabela 8.1. O SPSS lista o valor-p sob *Asymp. Sig.*, abreviatura de *significância assintótica*, onde "assintótico" se refere ao método de grandes amostras. A maioria dos *softwares* também relata uma estatística-teste alternativa, chamada de *estatística da máxima verossimilhança*, a qual fornece resultados similares. O Capítulo 15 apresenta essa estatística.

Interpretação dos graus de liberdade

Os gl em um teste qui-quadrado têm a seguinte interpretação: dados os totais marginais, as frequências das células em um bloco retangular do tamanho $(l - 1) \times (c - 1)$ dentro da tabela de contingência determinam as outras frequências da célula.

Para ilustrar, na Tabela 8.1, suponha que você conhece as duas frequências 573 e 516 na parte superior esquerda da tabela. Este é um bloco do tamanho 1×2, mostrado na Tabela 8.7. Então, dado os totais marginais, podemos determinar todas as frequências das outras células. Por exemplo, visto que 573 dos 959 Democratas são mulheres, os outros $959 - 573 = 386$ devem ser homens. Visto que 516 dos 991 Independentes são mulheres, os outros $991 - 516 = 475$ devem ser homens. Do mesmo modo, já que o total da linha das mulheres é 1511 e visto que as duas primeiras células contêm 1089 (isto é, 573 + 516) sujeitos, a célula remanescente deve ter $1511 - 1089 = 422$ observações. Disto e pelo fato de que a última coluna tem 821 observações, deve haver $821 - 422 = 399$ observações na segunda célula naquela coluna.

Uma vez que as frequências marginais estão fixadas em uma tabela de contingência, um bloco de somente $(l - 1) \times (c - 1)$ frequências de células é livre para variar, desde que estas frequências da célula determinam as remanescentes. O valor dos graus de liberdade é igual ao número de células neste bloco, assim $gl = (l - 1)(c - 1)$. Veremos outra forma de interpretar o gl no final da Seção 8.3.

Testes qui-quadrado e tratamento de categorias

No teste qui-quadrado, o valor da estatística-teste χ^2 não depende de qual é a variá-

☑ **Tabela 8.5** As cinco partes do teste qui-quadrado para independência

1. Suposições: Duas variáveis categóricas, amostragem aleatória, $f_e \geq 5$ em todas as células.
2. Hipóteses: H_0: As variáveis são independentes.
   H_a: As variáveis não são independentes.
3. Estatística-teste: $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$, onde $f_e = \frac{(\text{Total da linha})(\text{Total da coluna})}{\text{Tamanho amostral total}}$.
4. Valor-p: p = probabilidade da cauda direita acima do valor χ^2 observado, para a distribuição do qui-quadrado com gl = $(l - 1)(c - 1)$.
5. Conclusão: Informe o valor-p.
   Se necessário uma decisão, rejeite H_0 no nível α se $p \leq \alpha$.

☑ **Tabela 8.6** Saída do teste qui-quadrado para independência

SEX		PARTYID			
		democrata	independ	republ	Total
mulheres	Frequência	573	516	422	1511
	Frequência esperada	522,9	540,4	447,7	
homens	Frequência	386	475	399	1260
	Frequência esperada	436,1	450,6	373,3	
Total		959	991	821	2771

Estatística	Valor	gl	Sig. Assint.
Qui-quadrado de Pearson	16,202	2	0,000
Máxima verossimilhança	16,273	2	0,000

vel resposta e qual é a variável explicativa (se ambas). As etapas do teste e os resultados são idênticos em ambos os casos. Quando uma variável resposta é identificada e as distribuições condicionais da população são idênticas, elas são tidas como *homogêneas*. O teste qui-quadrado de independência é geralmente referido como um *teste de homogeneidade*. Por exemplo, a identificação partidária é a variável resposta e gênero é a explicativa, assim podemos considerar o teste qui-quadrado aplicado a estes dados como um teste de homogeneidade das distribuições condicionais da identificação partidária.

O teste qui-quadrado de independência trata as classificações como nominais. Isto é, χ^2 toma o mesmo valor se as linhas e colunas estão reordenadas de alguma forma. Se uma classificação é ordinal ou intervalar agrupada, o teste qui-quadrado não usa esta informação. Neste caso, geralmente é melhor aplicar métodos estatísticos mais fortes delineados para um nível mais alto de mensuração. A Seção 8.6 apresenta um teste de independência para variáveis ordinais.

8.3 RESÍDUOS: DETECTANDO O PADRÃO DE ASSOCIAÇÃO

O teste qui-quadrado de independência, como outros testes de significância, fornece informação limitada. Se o valor-p tem um tamanho moderado (por exemplo, $p > 0{,}10$), é plausível que as variáveis sejam independentes. Se o valor-p é muito pequeno, existe forte evidência de que as variáveis estejam associadas. O teste qui-quadrado não nos diz nada, entretanto, sobre a natureza ou a força da associação. O teste não indica se todas as células se desviam muito da independência ou talvez somente uma ou duas das células se desviam. As duas próximas seções introduzem métodos para aprender mais sobre a associação.

☑ **Tabela 8.7** Ilustração dos graus de liberdade; um bloco de $(I-1)(c-1)$ frequências de células conhecidas determina as outras

	Identificação partidária			
Gênero	Democrata	Independente	Republicano	Total
Mulheres	573	516	-	1511
Homens	-	-	-	1260
Total	959	991	821	2771

Análise dos resíduos

Uma comparação célula por célula de frequências observadas e esperadas revela a natureza da evidência sobre a associação. A diferença $(f_o - f_e)$ entre uma frequência observada e uma esperada de uma célula é chamada de um **resíduo**.

As frequências de identificação partidária e gênero são mostradas novamente a seguir, na Tabela 8.8. Para a primeira célula, o resíduo é igual a $573 - 522{,}9 = 50{,}1$. O resíduo é positivo quando, como nesta célula, a frequência observada f_o excede o valor f_e que a independência prevê. O resíduo é negativo quando a frequência observada é menor do que a independência prevê.

Como sabemos se o resíduo é grande o suficiente para indicar um desvio da independência que é improvável de ocorrer meramente por acaso? Uma forma padronizada do resíduo que se comporta como um escore-z fornece esta informação.

☑ **Resíduo padronizado**

O **resíduo padronizado** para uma célula é igual a:

$$z = \frac{f_o - f_e}{ep}$$

$$= \frac{f_o - f_e}{\sqrt{f_e(1 - \text{proporção da linha})\ (1 - \text{proporção da coluna})}}.$$

Aqui, ep representa o erro padrão de $f_o - f_e$, presumindo que H_0 é verdadeira. O resíduo padronizado é o número de erros padrão que $(f_o - f_e)$ está do valor 0, que é esperado quando H_0 é verdadeira.

O ep usa as proporções marginais para a linha e a coluna na qual a célula está. Quando H_0: independência for verdadeira, os resíduos padronizados têm uma distribuição normal padrão de amostras grandes. Eles flutuam em volta da média 0, com um desvio padrão de aproximadamente 1.

Usamos os resíduos padronizados de uma forma informal para descrever o padrão da associação entre as células. Um resíduo padronizado grande fornece evidência contra a independência naquela célula. Quando H_0 é verdadeira, existe somente em torno de 5% de chance de que qualquer resíduo padronizado exceda a 2 em valor absoluto. Quando inspecionamos várias células em uma tabela, alguns resíduos padronizados poderiam ser grandes apenas pela variação aleatória. Os valores abaixo de -3 ou acima de $+3$, entretanto, são evidências muito convincentes de um efeito verdadeiro naquela célula.

EXEMPLO 8.4 Resíduos padronizados para o gênero e a identificação partidária

A Tabela 8.8 exibe os resíduos padronizados para testar a independência entre gênero e afiliação partidária. Para a primeira célula, por exemplo, $f_o = 573$ e $f_e = 522{,}9$. A primeira linha e a primeira coluna das proporções marginais é igual a $1511/2771 = 0{,}545$ e $959/2771 = 0{,}346$. Substituindo na fórmula, o resíduo padronizado será:

$$z = \frac{f_o - f_e}{\sqrt{f_e(1 - \text{proporção da linha})\ (1 - \text{proporção da coluna})}}$$

$$= \frac{573 - 522{,}9}{\sqrt{[522{,}9(1 - 0{,}545)(1 - 0{,}346)]}} = 4{,}0.$$

Visto que o resíduo aleatorizado excede 3,0, esta célula tem mais observações do que esperaríamos se as variáveis fossem verdadeiramente independentes.

A Tabela 8.8 exibe resíduos positivos muito grandes para mulheres Democratas e homens Republicanos. Isso significa

que havia mais mulheres Democratas e homens Republicanos do que a hipótese de independência prevê. A tabela exibe resíduos negativos relativamente grandes para mulheres Republicanas e homens Democratas. Havia menos mulheres Republicanas e homens Democratas do que esperaríamos se a afiliação partidária fosse independente do gênero.

Para cada identificação partidária, a Tabela 8.8 contém somente um resíduo padronizado não redundante. O das mulheres é o oposto daqueles dos homens. As frequências observadas e as frequências esperadas têm os mesmos totais nas linhas e nas colunas. Portanto, em uma coluna dada, se $f_o > f_e$ em uma célula, o inverso deve acontecer na outra célula. As diferenças $f_o - f_e$ têm a mesma magnitude, mas sinal diferente nas duas células, indicando o mesmo padrão para os seus resíduos padronizados. ■

Junto com a estatística χ^2, a maioria dos *softwares* estatísticos pode fornecer os resíduos padronizados. Veja o apêndice do livro para mais detalhes.

O qui-quadrado e a diferença das proporções para tabelas 2 × 2

Como a Seção 7.2 (página 216) mostrou, as tabelas de contingência 2 × 2 geralmente comparam dois grupos em uma variável resposta binária. Os resultados poderiam ser, por exemplo, (sim, não) em questão de opinião. Para conveniência, rotulamos os dois resultados possíveis da variável binária por *sucesso* e *fracasso*.

Considere π_1 a representação da proporção de sucessos na população 1 e π_2 na população 2. Então, $(1 - \pi_1)$ e $(1 - \pi_2)$ são as proporções de fracassos. A Tabela 8.9 exibe a notação. As linhas são os grupos a serem comparados e as colunas são as categorias da resposta.

Se a variável resposta é estatisticamente independente das populações consideradas, então $\pi_1 = \pi_2$. A hipótese nula de independência corresponde à hipótese de *homogeneidade*, H_0: $\pi_1 = \pi_2$. Na verdade, o teste qui-quadrado de independência é equivalente a um teste para igualdade de duas proporções populacionais. A Seção 7.2 apresentou uma estatística-teste z para isto, baseada na divisão da diferença das proporções amostrais pelo seu erro padrão,

$$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{ep}.$$

A estatística qui-quadrado se relaciona a esta estatística z por $\chi^2 = z^2$.

A estatística qui-quadrado para tabelas 2 × 2 tem $gl = 1$. Seu valor-p da distribuição qui-quadrado é a mesma do valor-p para o teste bilateral com a estatística-teste z. Isto é devido à conexão direta entre a distribuição normal padrão e a distribuição do qui-quadrado com $gl = 1$: elevar ao quadrado os escores-z com certas probabilidades de duas caudas gera escores do qui-quadrado com $gl = 1$ tendo as mesmas probabilidades na cauda direita. Por exemplo, $z = 1{,}96$ é o escore-z com a probabilidade nas duas caudas de 0,05. O quadrado disso, $(1{,}96)^2 = 3{,}84$, é o escore do qui-quadrado para $gl = 1$ com o valor-p de 0,05. (Você pode verificar isso na Tabela C.)

☑ **Tabela 8.8** Resíduos padronizados (nos parênteses) para testar a independência entre afiliação partidária e gênero

	Identificação partidária		
Gênero	Democrata	Independente	Republicano
Mulheres	573(4,0)	516(−1,9)	422 (−2,1)
Homens	386(−4,0)	475(1,9)	399 (2,1)

☑ **Tabela 8.9** Tabela 2 × 2 para comparar dois grupos em uma variável resposta binária

	Proporção de cada resposta		
Grupo	Sucesso	Fracasso	Total
1	π_1	$1 - \pi_1$	1,0
2	π_2	$1 - \pi_2$	1,0

EXEMPLO 8.5 Os papéis de homens e mulheres

A Tabela 8.10 resume as respostas das Pesquisas Sociais Gerais de 1977 e de 2006 para a declaração ("FEFAM"): "Seria melhor para todos os envolvidos que o homem fosse o provedor fora de casa e a mulher tomasse conta da casa e da família". Você pode verificar que a proporção amostral concordando com a declaração foi de $\hat{\pi}_1 = 0{,}658$ em 1977 e $\hat{\pi}_2 = 0{,}358$ em 2006, o *ep* para o teste comparando-as é igual a 0,0171 e a estatística-teste z para H_0: $\hat{\pi}_1 = \hat{\pi}_2$ é $z = (0{,}658 - 0{,}358)/0{,}0171 = 17{,}54$. Você também pode verificar que a estatística qui-quadrado para esta tabela é $\chi^2 = 307{,}6$. Isto é igual ao quadrado da estatística-teste z. Ambas as estatísticas mostram evidência extremamente forte contra a hipótese nula de proporções populacionais iguais. ■

Resíduos padronizados para tabelas 2 × 2

Vamos continuar com o teste para a Tabela 8.10 com uma análise de resíduos. A Tabela 8.10 também mostra os resíduos padronizados. Aqueles na primeira coluna sugerem que mais sujeitos concordavam com a afirmação em 1977 e menos concordavam em 2006 do que esperaríamos se as opiniões fossem independentes no ano da pesquisa. Observe que *cada* resíduo padronizado é igual a +17,5 ou −17,5. O valor absoluto do resíduo padronizado é de 17,5 em cada célula.

Para testes qui-quadrado com tabelas 2 × 2, $gl = 1$. Isto significa que existe somente uma informação sobre se há uma associação. Uma vez que encontramos o resíduo padronizado para uma célula, outros resíduos padronizados na tabela têm o mesmo valor absoluto. Na verdade, nas tabelas 2 × 2, cada resíduo padronizado é igual à estatística-teste z (ou o seu negativo) para comparar duas proporções. O quadrado de cada resíduo padronizado é igual à estatística-teste χ^2.

O qui-quadrado necessário para tabelas maiores do que 2 × 2

Para uma tabela 2 × 2, por que deveríamos executar um teste z se podemos obter o mesmo resultado com o qui-quadrado? Uma vantagem do teste z é que ele também aplica hipóteses alternativas unilaterais, como H_a: $\pi_1 > \pi_2$. A direção do

☑ **Tabela 8.10** Respostas da PSG à declaração "Seria melhor para todos os envolvidos que o homem fosse o provedor fora de casa e a mulher tomasse conta da casa e da família", com os resíduos padronizados em parênteses

Ano	Concorda	Discorda	Total
1977	989(17,5)	514(−17,5)	1503
2006	704(−17,5)	1264(17,5)	1968

efeito é perdida elevando z ao quadrado e usando o χ^2.

Por que necessitamos da estatística χ^2? A razão é que uma estatística z pode somente comparar uma única estimativa a um único valor H_0. Exemplos são a estatística z para comparar uma proporção amostral a uma sob H_0 como 0,5 ou uma diferença de proporções amostrais ao valor H_0 de 0 para $\pi_2 - \pi_1$. Quando uma tabela for maior do que 2×2 e, dessa forma, $gl > 1$, necessitamos mais do que um parâmetro de diferença para descrever a associação. Por exemplo, suponha que a Tabela 8.10 tivesse três linhas, para três anos de dados. Então, H_0: independência corresponde a $\pi_1 = \pi_2 = \pi_3$, onde π_i é a proporção da população concordando com a afirmação no ano i. Os parâmetros de comparação são ($\pi_1 - \pi_2$), $(\pi_1 - \pi_3)$ e $(\pi_2 - \pi_3)$. Poderíamos usar a estatística z para cada comparação, mas não uma única estatística z para todo o teste de independência.

Podemos interpretar o valor gl em um teste qui-quadrado como o número de parâmetros necessários para determinar todas as comparações para descrever a tabela de contingência. Por exemplo, em uma tabela de contingência 3×2, para comparar três anos em uma resposta de opinião binária, $gl = 2$. Isto significa que precisamos comparar somente dois parâmetros para poder descobrir a terceira comparação. Por exemplo, se conhecemos $(\pi_1 - \pi_2)$ e $(\pi_1 - \pi_3)$, então:

$$(\pi_2 - \pi_3) = (\pi_1 - \pi_3) - (\pi_1 - \pi_2).$$

8.4 MEDINDO A ASSOCIAÇÃO EM UMA TABELA DE CONTINGÊNCIA

As principais perguntas normalmente feitas na análise de uma tabela de contingência são as seguintes:

- *Existe uma associação*? O teste qui-quadrado de independência trata desta pergunta. Quanto menor o valor-p mais forte a evidência de associação.
- *Como os dados diferem do que a independência prevê*? Os resíduos padronizados destacam as células que apresentam valores maiores ou menores do que o esperado sob a hipótese de independência.
- *Quão forte é a associação?* Para resumir isto, usamos uma estatística como uma diferença de proporções, formando um intervalo de confiança para estimar a força da associação na população.

Analisar a *força* da associação revela se a associação é importante ou se ela é estatisticamente significativa, mas sem significância prática. Esta seção apresenta duas formas de medir a força da associação para tabelas de contingência.

Medidas de associação

> ☑ **Medidas de associação**
>
> Uma **medida de associação** é uma estatística ou um parâmetro que resume a força da dependência entre duas variáveis.

Vamos, em primeiro lugar, considerar o que quer dizer uma associação *forte versus fraca*. A Tabela 8.11 mostra duas tabelas de contingências hipotéticas relacionando raça à opinião sobre a permissão de união civil para casais do mesmo sexo. O Caso A, que exibe independência estatística, representa a associação mais fraca possível. Negros e brancos têm 60% a favor e 40% contra uniões civis. A opinião não está associada à raça. Em contraposição, o Caso B exibe a associação mais forte possível. Todos os brancos são a favor das uniões civis, enquanto todos os negros são contrários. Nesta tabela, a opinião é completamente dependente da raça. Para estes sujeitos, se conhecemos a sua raça, nós sabemos a sua opinião.

Uma medida de associação descreve o quão similar uma tabela é às tabelas representando as associações mais fortes e mais fracas. Ela assume um intervalo de valores de um extremo a outro conforme os dados variam de uma associação mais fraca a uma mais forte.

☑ **Tabela 8.11** Classificação cruzada da opinião sobre uniões civis do mesmo sexo por raça mostrando (A) Nenhuma associação, (B) Associação máxima

Caso A		Opinião			Caso B	Opinião		
	Raça	A Favor	Contra	Total		A Favor	Contra	Total
	Branca	360	240	600		600	0	600
	Negra	240	160	400		0	400	400
	Total	600	400	1000		600	400	1000

Diferença de proporções

Como foi discutido nas Seções 7.2 (página 216) e 8.3, muitas tabelas 2×2 comparam dois grupos em uma variável binária. Em tais casos, uma medida de associação de fácil interpretação é a diferença entre as proporções para uma categoria de resposta dada. Por exemplo, poderíamos medir a diferença entre as proporções de brancos e negros que são a favor da permissão de uniões civis entre o mesmo sexo. Para a Tabela 8.11, Caso A, esta diferença é:

$$\frac{360}{600} - \frac{240}{400} = 0{,}60 - 0{,}60 = 0{,}0.$$

A diferença das proporções populacionais é 0 sempre que as distribuições condicionais sejam idênticas, isto é, quando as variáveis são independentes. A diferença é 1 ou -1 para a associação mais forte possível. Para a Tabela 8.11, Caso B, por exemplo, a diferença é:

$$\frac{600}{600} - \frac{0}{400} = 1{,}0,$$

o valor absoluto máximo possível para a diferença.

Esta medida está entre -1 e $+1$. Na prática não esperamos que os dados assumam valores extremos, mas, quanto *maior a associação, maior o valor absoluto da diferença das proporções*. As tabelas de contingência seguintes ilustram o aumento nesta medida com o aumento do grau da associação:

Frequências das Células:	25 25 25 25	30 20 20 30	35 15 15 35	40 10 10 40	45 5 5 45	50 0 0 50
Diferença das Proporções:	0	0,2	0,4	0,6	0,8	1,0

Para a segunda tabela, por exemplo, a proporção que está na primeira coluna é igual a $30/(30 + 20) = 0{,}60$ na linha 1 e $20/(20 + 30) = 0{,}40$ na linha 2, para uma diferença de $0{,}60 - 0{,}40 = 0{,}20$.

O qui-quadrado não mensura a associação

Um valor alto para o χ^2 no teste de independência sugere que as variáveis estão associadas. Isto não implica, entretanto, que as variáveis tenham uma associação forte. Essa estatística resume quão próximas as frequências observadas estão das frequências esperadas se as variáveis fossem independentes. Ela simplesmente indica quanta evidência existe de que as variáveis sejam dependentes, não quão forte é a dependência. Para uma associação dada, valores altos do χ^2 ocorrem para tamanhos de amostra

maiores. Assim, como com qualquer teste de significância, valores grandes da estatística-teste podem ocorrer com efeitos fracos, se o tamanho da amostra for grande.

Por exemplo, considere os casos hipotéticos da Tabela 8.12. A associação em cada tabela é muito fraca – a distribuição condicional para os brancos quanto à opinião (49% a favor, 51% contra) é aproximadamente idêntica à distribuição condicional para os negros (51% a favor, 49% contra). Todas as três tabelas mostram exatamente o mesmo grau de associação, com a diferença entre as proporções de negros e brancos que são a favor da legalização das uniões civis do mesmo sexo sendo $0,51 - 0,49 = 0,02$ em cada tabela.

Para o tamanho amostral de 200, no Caso A, $\chi^2 = 0,08$, que tem um valor-p = 0,78. Para um tamanho amostral de 400, no Caso B, $\chi^2 = 0,16$, com um valor-p = 0,69. Assim, quando as frequências das células dobram, o χ^2 dobra. Da mesma forma, para um tamanho de amostra de 20000 (100 vezes o valor de $n = 200$), no Caso C, $\chi^2 = 8,0$ (100 vezes o valor do $\chi^2 = 0,08$) com um valor-$p = 0,005$.

Em resumo, para um percentual fixo atribuído às células de uma tabela de contingência, o valor do χ^2 é diretamente proporcional ao tamanho da amostra – valores maiores ocorrem com tamanhos da amostra maiores. Como ocorre com outras estatísticas-teste, quanto maior a estatística χ^2 menor será o seu valor-p e mais forte a evidência contra a hipótese nula. Entretanto, um valor-p pequeno pode resultar de uma associação fraca quando o tamanho da amostra é grande, como mostra o Caso C.

A razão de chances*

A diferença das proporções é de fácil interpretação. Muitas outras medidas são também relatadas por um *software* estatístico. Esta subseção apresenta o mais importante para a análise de dados categóricos, a razão de chances.

Para uma variável resposta binária, lembre que usamos *sucesso* para representar o resultado do interesse e *fracasso* para o outro resultado. A **chance** de sucesso é definida como:

$$\text{Chance} = \frac{\text{Probabilidade de sucesso}}{\text{Probabilidade de fracasso}}.$$

Se a probabilidade de sucesso = 0,75, então a probabilidade de fracasso é igual $1 - 0,75 = 0,25$ e a chance de sucesso = 0,75/0,25 = 3,0. Se P(sucesso) = 0,50, então a chance = 0,50/0,50 = 1,0. Se P(sucesso) = 0,25, então a chance = 0,25/0,75 = 1/3. A chance é um valor não negativo e maior do que 1,0 quando um sucesso é mais provável do que um fracasso. Quando a chance = 3,0, um sucesso é três vezes mais provável do que um fracasso; esperamos aproximadamente três sucessos para cada fracasso. Quando a chance = 1/3, um fracasso é três vezes mais provável que um sucesso; esperamos aproximadamente um sucesso a cada três fracassos.

A probabilidade de um resultado está relacionada à chance desse mesmo resultado por:

$$\text{Probabilidade} = \frac{\text{Chance}}{\text{Chance} + 1}$$

Por exemplo, quando a chance = 3, a probabilidade = 3/(3 + 1) = 0,75.

O quociente de chances de duas linhas de uma tabela 2 × 2 é chamada de **razão de chances**. Por exemplo, se a chance = 4,5, na linha 1 e a chance = 3,0, na linha 2, então a razão de chances é igual a 4,5/3,0 = 1,5. A chance de sucesso na linha 1 é, então, igual a 1,5 vezes a chance de sucesso da linha 2. Representamos a razão de chances pela letra grega θ (teta).

☑ **Tabela 8.12** Classificação cruzada da opinião sobre uniões civis do mesmo sexo, por raça, mostrando associações fracas, mas idênticas

	A			B			C		
	Sim	Não	Total	Sim	Não	Total	Sim	Não	Total
Branco	49	51	100	98	102	200	4900	5100	10000
Negro	51	49	100	102	98	200	5100	4900	10000
	100	100	200	200	200	400	10000	10000	20000
	$\chi^2 = 0,08$ Valor-p = 0,78			$\chi^2 = 0,16$ Valor-p = 0,69			$\chi^2 = 8,0$ Valor-p = 0,005		

EXEMPLO 8.6 Raça das vítimas de assassinato e criminosos

Para assassinatos nos Estados Unidos em 2005 tendo uma única vítima e um único criminoso, a Tabela 8.13 faz uma classificação cruzada entre a raça da vítima e a do criminoso. Tratamos a raça da vítima como a variável resposta. Para criminosos brancos, a proporção de vítimas também brancas é 3150/3380 = 0,932, e a proporção negra é 230/3380 = 0,068. A chance de uma vítima ser branca é igual a 0,932/0,068 = 13,7. Isto é igual a (3150/3380)/(230/3380) = 3150/230. Assim, podemos calcular a chance pela razão das frequências nas duas células na linha 1, sem convertê-las em proporções.

O valor 13,7 significa que, para criminosos brancos, ocorreram 13,7 vítimas brancas para cada 1 vítima negra. Para criminosos negros, a chance de uma vítima ser branca é igual a 516/2984 = 0,173. Isto significa que ocorreu 0,173 vítimas brancas para cada 1 vítima negra. De forma equivalente, visto que 2984/516 = 1/0,173 = 5,8, criminosos negros tinham 5,8 vítimas negras para cada vítima branca.

Para a Tabela 8.13, a razão de chances é igual a:

$$\theta = \frac{\text{Chances para criminosos brancos}}{\text{Chances para criminosos negros}} = \frac{13,7}{0,173} = 79,2$$

Para os criminosos brancos, a chance de uma vítima ser branca foram de aproximadamente 79 vezes a chance de uma ser vítima branca para criminosos negros. ■

Em resumo:

> ☑ **Chance e razão de chances**
>
> A **chance** estimada para uma resposta binária é igual ao número de sucessos dividido pelo número de fracassos.
>
> A **razão de chances** é uma medida da associação para tabelas de contingência 2 × 2 que é igual à chance na linha 1 dividida pela chance na linha 2.

Propriedades da razão das chances*

Na Tabela 8.13, suponha que tratamos da raça do criminoso em vez da raça da vítima como a variável resposta. Quando as vítimas eram brancas, a chance de que a raça do criminoso fosse branca era 3150/516 = 6,10. Quando a vítima era ne-

☑ **Tabela 8.13** Classificação cruzada entre as raças da vítima e do criminoso

Raça do criminoso	Raça da vítima		
	Branca	Negra	Total
Branca	3150	230	3380
Negra	516	2984	3500

Fonte: www.fbi.gov

gra, a chance de que a raça do criminoso fosse branca era 230/2984 = 0,077. A razão de chances é igual a 6,10/0,077 = 79,2. Para cada escolha da variável resposta, a razão de chances é 79,2. Na verdade:

- A razão de chances assume o mesmo valor a despeito da escolha da variável resposta. Visto que a razão de chances trata as variáveis simetricamente, a razão das chances é uma medida natural quando não existe distinção óbvia entre as variáveis, como quando ambas são variáveis resposta.
- A razão de chances θ é igual à razão dos produtos das frequências das células opostas em diagonal.
  Para a Tabela 8.13, por exemplo,

$$\theta = \frac{(3150 \times 2984)}{(230 \times 516)} = 79{,}2.$$

Em virtude dessa propriedade, a razão das chances é também chamada de **razão do produto cruzado**.

- A razão de chances pode ser qualquer número não negativo.
- Quando as probabilidades de sucesso são idênticas nas duas linhas de uma tabela 2 × 2 (isto é, $\pi_1 = \pi_2$), então $\theta = 1$.

Quando $\pi_1 = \pi_2$, as chances são também iguais. As chances de sucesso não dependem do nível da linha da tabela, e as variáveis são, então, independentes, com $\theta = 1$. O valor $\theta = 1$ para independência serve como uma linha-base de comparação. A razão de chances em cada lado do valor 1 reflete certos tipos de associações.

- Quando $\theta > 1$, a chance de sucesso é *maior* na linha 1 do que na linha 2.

Por exemplo, quando $\theta = 4$, a chance de sucesso na linha 1 é quatro vezes a chance de sucesso na linha 2.

- Quando $\theta < 1$, a chance de sucesso é *menor* na linha 1 do que na linha 2.
- Valores de θ mais distantes de 1,0 em uma determinada direção representam associações mais fortes.

Uma razão de chances de 4 está mais distante da independência do que uma de 2 e uma razão de chances de 0,25 está mais distante da independência do que uma de 0,50.

- Dois valores para θ representam a mesma força da associação, mas em direções opostas, quando um valor é o recíproco do outro.

Por exemplo, $\theta = 4{,}0$ e $\theta = 1/4 = 0{,}25$ representam a mesma força da associação. Quando $\theta = 0{,}25$, a chance de sucesso na linha 1 é 0,25 vezes a chance de sucesso da linha 2. De forma equivalente, a chance de sucesso na linha 2 é 1/0,25 = 4,0 vezes a chance de sucesso da linha 1. Quando a ordem das linhas é invertida ou a ordem das colunas é invertida, o novo valor de θ é o recíproco do valor original. Essa disposição das linhas ou colunas é usualmente arbitrária, assim se obtemos 4,0 ou 0,25 para a razão de chances é uma simples questão de como rotulamos as linhas e colunas.

Na interpretação da razão de chances, tenha cuidado para não interpretá-la erroneamente como uma razão de probabilidades. Uma razão de chances de 79,2 *não* significa que π_1 é 79,2 vezes π_2. Ao contrário, $\theta = 79{,}2$ significa que a *chance* na linha 1 é igual a 79,2 vezes a chance na linha 2. A razão de chances é um quociente entre duas chances, não uma razão entre duas probabilidades. Isto é:

$$\theta = \frac{\text{Chance na linha 1}}{\text{Chance na linha 2}} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}, \quad \text{não} \quad \frac{\pi_1}{\pi_2}.$$

A razão π_1/π_2 é, ela própria, uma medida útil. A Seção 7.1 (página 212) introduziu essa medida que é geralmente denominada de **risco relativo**.

A distribuição amostral da razão de chances $\hat{\theta}$ é altamente assimétrica a não ser que o tamanho da amostra seja extremamente grande, neste caso a distribuição é aproximadamente normal. Veja o Exercício 8.45 para um método da construção de intervalos de confiança para razões de chances.

Razões de chances para tabelas $I \times c$*

Para tabelas de contingência com mais do que duas linhas ou mais do que duas colunas, a razão de chances descreve padrões em qualquer subtabela 2 × 2. Ilustramos isso utilizando os dados da PSG da identificação partidária *versus* raça, mostrado na Tabela 8.14.

Considere, em primeiro lugar, a subtabela 2 × 2 formada das duas primeiras colunas. A razão de chances amostral é igual a (250 × 783)/(106 × 640) = 2,89. A chance de que a resposta de um negro tenha sido Democrata em vez de Independente é igual a 2,89 vezes a chance de um branco. Daqueles sujeitos que responderam Democrata ou Independente, os negros tinham maior probabilidade do que os brancos de responder Democrata.

A razão de chances amostral para as duas últimas colunas desta tabela é igual a (106 × 775)/(17 × 783) = 6,17. A chance de que a resposta de um negro fosse Independente em vez de Republicano é igual a 6,2 vezes a chance de um branco. Daqueles sujeitos que responderam Independente ou Republicano, os negros tinham maior probabilidade do que os brancos de responder Independente.

Finalmente, para a subtabela 2 × 2 formada pela primeira e última coluna, a razão de chances amostral é igual a (250 × 775)/(17 × 640) = 17,81. A chance de que a resposta de um negro fosse Democrata em vez de Republicano é igual a 17,8 vezes a chance de um branco. Daqueles sujeitos que responderam Democrata ou Republicano, os negros tinham uma probabilidade maior do que os brancos de responder Democrata. Este é um efeito muito forte, bem distante do valor da razão de chances de independência de 1,0.

O valor da razão de chances de 17,8 para a primeira e última coluna é igual a (2,89)(6,17), o produto das outras duas razões de chances. Para tabelas 2 × 3, um $gl = 2$ significa que existem somente duas informações sobre a associação. As duas razões das chances determinam a terceira.

Medidas resumo de associação para tabelas $I \times c$*

Em vez de estudar a associação nas subtabelas 2 × 2, é possível determinar a associação em toda a tabela por meio de um único número. Uma forma de fazer isto resume quão bem podemos prever o valor em uma variável com base no conhecimento do valor da outra variável. Por exemplo, a identificação partidária e a raça estão altamente associadas se a raça for um bom previsor da identificação partidária, isto é, se sabemos a raça, podemos fazer previsões muito melhores sobre a identificação partidária das pessoas do que se não soubéssemos.

Para variáveis quantitativas, a *correlação* é uma medida resumo. Estudare-

☑ **Tabela 8.14** Dados da PSG de 2004 sobre a identificação partidária e raça

	Identificação partidária		
Raça	Democrata	Independente	Republicano
Negra	250	106	17
Branca	640	783	775

mos uma medida resumo similar desse tipo para variáveis ordinais (denominada *gama*) na próxima seção. Esta medida descreve uma tendência geral dos dados. Para variáveis nominais, quando l ou c excede 2, geralmente é uma simplificação excessiva descrever a tabela com uma única medida da associação. Neste caso, existem muitos padrões possíveis de associação para que uma tabela $l \times c$ seja descrita por um único número. As medidas nominais baseadas no poder de previsão (chamadas de *tau* e *lambda*) e a *gama* para dados ordinais foram definidas em 1954 por dois proeminentes cientistas estatísticos sociais, Leo Goodman e William Kruskal. A maioria dos *softwares* para analisar tabelas de contingência apresenta essas medidas e várias outras. Algumas medidas nominais, como o coeficiente de contingência e o V de Cramer, são difíceis de interpretar (exceto valores altos representando associação mais forte) e, em nossa opinião, não são especialmente úteis.

Não apresentamos medidas nominais de resumo neste livro. Acreditamos que você irá entender melhor a associação fazendo comparações dos percentuais de distribuições condicionais, visualizando o padrão dos resíduos padronizados nas células da tabela, construindo razões de chances em subtabelas 2×2 e construindo modelos tais como os apresentados no Capítulo 15. Esses métodos tornam-se ainda mais preferíveis como medidas de resumo de uma associação quando a análise é multivariada em vez de bivariada.

8.5 ASSOCIAÇÃO ENTRE VARIÁVEIS ORDINAIS*

Agora, voltaremos nossa atenção a outras análises de tabelas de contingência que se aplicam quando as variáveis são ordinais. As categorias de variáveis ordinais são ordenadas. Análises estatísticas para dados ordinais levam essa ordenação em consideração. Esta seção introduz uma medida ordinal popular de associação, e a Seção 8.6 apresenta métodos relacionados de inferência.

EXEMPLOS 8.7 Quão fortemente associados estão renda e felicidade?

A Tabela 8.15 é uma tabela de contingência com variáveis ordinais. Estes dados, da PSG de 2004, se referem à relação entre renda familiar ("FINRELA") e felicidade ("HAPPY"). Esta tabela mostra resultados para norte-americanos negros, e o Exercício 8.13 analisa os dados para norte-americanos brancos.

Vamos inicialmente ter uma ideia dos dados estudando as distribuições condicionais da variável felicidade. A Tabela 8.15 mostra as distribuições em parênteses. Por exemplo, a distribuição condicional (24%, 54%, 22%) exibe os percentuais nas categorias da felicidade para sujeitos com renda familiar abaixo da média. Somente 22% estão muito felizes, enquanto 36% dos sujeitos no grupo do nível mais alto de renda estão muito felizes. Inversamente, um percentual mais baixo (9%) do grupo de alta renda não está tão feliz comparado com o grupo de renda mais baixa (24%). A razão de chances

☑ **Tabela 8.15** Rendimento familiar e felicidade para uma amostra da PSG

	Felicidade			
Renda familiar	Pouco feliz	Feliz	Muito feliz	Total
Abaixo da média	16 (24%)	36 (54%)	15 (22%)	67(100,0%)
Na média	11 (16%)	36 (53%)	21 (31%)	68(100,0%)
Acima da média	2 (9%)	12 (55%)	8 (36%)	22(100,0%)
Total	29	84	44	157

para as quatro células do canto é $(16 \times 8)/(15 \times 2) = 4{,}3$. Parece que os sujeitos com maior renda tendem a ser mais felizes. ■

Os dados ordinais exibem dois tipos primários de associação entre as variáveis x e y – *positiva* e *negativa*. Uma associação positiva ocorre quando os sujeitos no final superior da escala em x tendem também a estar no final superior da escala em y e aqueles que estão na parte inferior de x tendem a estar na parte inferior de y. Por exemplo, existe uma associação positiva entre renda e felicidade se aqueles com baixa renda tendem a ser pouco felizes e os com altas rendas tendem a ter a ser maiz felizes. A associação negativa ocorre quando os sujeitos classificados na parte superior de x tendem a ser classificados na parte inferior de y e aqueles classificados na parte inferior de x tendem a estar na parte superior de y. Por exemplo, uma associação negativa pode existir entre fundamentalismo religioso e tolerância em relação à homossexualidade – quanto mais fundamentalista nas crenças religiosas, menos tolerante em relação à homossexualidade.

Concordância e discordância

Muitas medidas ordinais de associação são baseadas na informação sobre a associação fornecida por todos os pares de informação.

☑ **Par concordante, par discordante**

Um par de observações é **concordante** se o sujeito que tem um valor alto em uma variável apresenta também um valor alto na outra variável.
Um par de observações é **discordante** se o sujeito que tem um valor alto em uma variável apresenta um valor baixo na outra.

Na Tabela 8.15, consideramos *Pouco feliz* (PF) como a parte inferior e *Muito feliz* (MF) como a parte superior da escala y = felicidade, e *Abaixo da média* como inferior e *Acima da média* como superior na escala x = renda familiar. Por convenção, construímos tabelas de contingência para variáveis ordinais de modo que a última categoria da variável linha esteja na primeira e a última categoria da variável coluna seja a primeira. (Não existe um padrão, entretanto, e outros livros ou *softwares* podem usar uma convenção diferente.)

Considere um par de sujeitos, um dos quais é classificado (abaixo da média, pouco feliz) e o outro é classificado como (na média, feliz [F]). O primeiro sujeito é um dos 16 classificados na parte superior esquerda da célula da Tabela 8.15, e o segundo é um dos 36 classificados na célula do meio. Esse par de sujeitos é concordante, visto que o segundo sujeito está acima do primeiro tanto em felcidade quanto em renda. O sujeito que está acima em uma variável também está acima na outra. Agora, cada um dos 16 sujeitos classificados (abaixo da média, pouco feliz) pode formar um par com cada um dos 36 sujeitos classificados (na média, feliz). Portanto, existem $16 \times 36 = 576$ pares concordantes destas duas células.

Em contraposição, cada um dos 36 sujeitos na célula (abaixo da média, feliz) forma um par discordante quando equiparado com cada um dos 11 sujeitos na célula (na média, pouco feliz). Os 36 sujeitos têm uma renda mais baixa do que os outros 11 sujeitos, mas eles têm muita felicidade. Todos os $36 \times 11 = 396$ desses pares de sujeitos são discordantes.

Pares concordantes de observações fornecem evidência de associação positiva, desde que, para tal par, o sujeito que está acima em uma variável também esteja acima na outra. Por outro lado, quanto mais predominantes forem os pares discordantes, maior a evidência de uma associação negativa.

☑ **Notação para o números de pares concordantes e discordantes**

Considere C a representação do número total de pares concordantes e D a do número total de pares discordantes.

Uma regra geral para encontrar o número de pares concordantes C é esta: inicie no canto da tabela que apresenta as menores categorias para cada variável (a célula na linha 1 e coluna 1 para a Tabela 8.15). Multiplique a frequência daquela célula pela maior frequência em cada célula em ambas as variáveis (aquelas células abaixo e à direita na Tabela 8.15). Da mesma forma, para cada uma das demais células, multiplique a frequência da célula pelas maiores frequências em ambas as variáveis. (Para as células na linha ou na coluna no maior nível da variável, como a linha *Acima da média* ou a coluna *Muito feliz* na Tabela 8.15, nenhuma observação é maior em ambas as variáveis.) O número de pares concordantes é a soma desses produtos.

Na Tabela 8.15, os 16 sujeitos na primeira célula são concordantes quando equiparados com os (36 + 21 + 12 + 8) sujeitos abaixo e à esquerda que estão no nível mais alto em cada variável. Da mesma forma, os 36 sujeitos na segunda célula na primeira linha são concordantes quando equiparados com os (21 + 8) sujeitos que estão no nível mais alto em cada variável e assim por diante. Portanto:

$$C = 16(36 + 21 + 12 + 8) + 36(21 + 8) + 11(12 + 8) + 36(8) = 2784.$$

A Tabela 8.16 exibe o cálculo do número total de pares concordantes.

Para encontrar o número total de pares discordantes D, inicie no canto da tabela que apresenta o nível mais alto de uma variável e o mais baixo da outra. Por exemplo, os 15 sujeitos na célula (abaixo da média, muito feliz) formam pares discordantes quando emparelhados com os (11 + 36 + 2 + 12) sujeitos abaixo e à esquerda da tabela que estão no nível mais alto de renda, mas mais baixo em felicidade. Multiplique a frequência de cada célula pelas frequências de todas as células que estão no nível mais alto de renda, mas mais baixas em felicidade. O número total de pares discordantes é:

$$D = 15(11 + 36 + 2 + 12) + 21(2 + 12) + 36(11 + 2) + 36(2) = 1749.$$

A Tabela 8.17 exibe o cálculo do número de pares discordantes.

Em resumo, a Tabela 8.15 tem $C = 2784$ e $D = 1749$. Mais pares mostram evidência de uma associação positiva (isto é, pares concordantes) do que mostram evidência de uma associação negativa (pares discordantes).

Gama

Uma diferença positiva para $C - D$ ocorre quando $C > D$. Isso indica uma associação positiva. Uma diferença negativa para $C - D$ reflete uma associação negativa.

Amostras maiores têm números maiores de pares com, normalmente, diferenças absolutas maiores em $C - D$. Portanto, padronizamos esta diferença para tornar mais fácil a interpretação. Para fazer isso, dividimos $C - D$ pelo número total de pares que são ou concordantes ou discordantes, $C + D$. Isto fornece a medida de associação chamada de **gama**. Sua fórmula amostral é:

☑ **Tabela 8.16** Ilustração do cálculo do número de pares concordantes, C

	PF	F	MF
Abaixo da média	16		
Na média		36	21
Acima da média		12	8

$C = 16(36+21+12+8)$ + 36(21 + 8) + 11(12 + 8) + 36(8) = 2784

	PF	F	MF
Abaixo da média			15
Na média	11	36	
Acima da média	2	12	

$D = 15(11+36+2+12)$ + 21(2 + 12) + 36(11 + 2) + 36(2) = 1749

$$\hat{\gamma} = \frac{C - D}{C + D}.$$

Aqui estão algumas propriedades de gama:

- O valor de gama está entre −1 e +1.
- O sinal de gama indica se a associação é positiva ou negativa.
- Quanto maior o valor absoluto de gama, maior a associação.

A tabela para a qual gama é igual a 0,60 ou −0,60 exibe uma associação mais forte do que uma para a qual gama é igual a 0,30 ou −0,30, por exemplo. O valor +1 representa a associação positiva mais forte. Isso ocorre quando não existem pares discordantes ($D = 0$), assim todos os pares revelam uma associação positiva. Gama é igual −1 quando $C = 0$, assim todos os pares revelam uma associação negativa. Gama é igual a 0 quando $C = D$.

Para a Tabela 8.15, $C = 2784$ e $D = 1749$, portanto:

$$\hat{\gamma} = \frac{2784 - 1749}{2784 + 1749} = 0{,}228.$$

Essa amostra exibe uma associação positiva entre a renda familiar e felicidade. Quanto maior a renda familiar, maior a tendência de felicidade. Entretanto, o valor amostral está mais próximo de 0 do que de 1, de modo que a associação é relativamente fraca.

O cálculo de gama é muito confuso. A maioria dos *softwares* estatísticos pode encontrar gama para você.

Gama é a diferença entre duas proporções ordinais

Outra interpretação para a magnitude de gama segue da expressão:

$$\hat{\gamma} = \frac{C - D}{C + D} = \frac{C}{C + D} - \frac{D}{C + D}.$$

Agora, $(C + D)$ é o número total de pares que são concordantes e discordantes. A razão $C/(C + D)$ é a proporção desses pares que são concordantes, $D/(C + D)$ é a proporção dos pares que são discordantes e $\hat{\gamma}$ é a diferença entre as duas proporções.

Por exemplo, suponha que $\hat{\gamma} = 0{,}60$. Então, visto que 0,80 e 0,20 são as duas proporções que somam 1 e têm uma diferença de 0,80 − 0,20 = 0,60, 80% dos pares são concordantes e 20% dos pares são discordantes. Da mesma forma, $\hat{\gamma} = -0{,}333$ indica que 1/3 dos pares são concordantes e 2/3 dos pares são discordantes, desde que 1/3 + 2/3 = 1 e 1/3 − 2/3 = −0,333.

Para a Tabela 8.15, dentre os 2784 + 1749 = 4533 pares que são concordantes ou discordantes, a proporção 2784/4533 = 0,614 são concordantes e a proporção 1749/4533 = 0,386 são discordantes; $\hat{\gamma}$ = 0,228 é a diferença entre essas proporções.

Propriedades comuns das medidas ordinais

Gama é uma das várias medidas de associação. As outras são **tau-*b*** e **tau-*c* de Kendall**, **rho-*b*** e **rho-*c* de Spearman** e o ***d* de Somers**. Todas estas medidas são similares em seus propósitos básicos e característi-

cas. Por falta de espaço, não definiremos essas medidas, mas iremos listar algumas propriedades comuns. Estas propriedades também se mantêm para a *correlação* para variáveis quantitativas que foram introduzidas na Seção 3.5 (página 73) e serão extensivamente usadas no próximo capítulo.

- Medidas ordinais de associação assumem valores entre -1 e $+1$. O sinal nos diz se a associação é positiva ou negativa.
- Se as variáveis são estatisticamente independentes, então os valores da população de medidas ordinais são iguais a 0.
- Quanto maior a associação, maior o valor absoluto da medida. Valores de 1,0 e $-1,0$ representam associações mais fortes.
- Com a exceção do d de Somers, as medidas originais das associações mencionadas acima não distinguem entre variáveis respostas e explicativas. Elas assumem o mesmo valor tanto para y sendo a variável resposta quanto sendo explicativa.

Até agora, discutimos o uso de medidas ordinais somente para descrever, ou seja, como medidas descritivas. A próxima seção apresenta a inferência estatística, a saber, intervalos de confiança e testes para dados ordinais.

8.6 INFERÊNCIA PARA ASSOCIAÇÕES ORDINAIS*

O teste qui-quadrado para verificar se duas variáveis categóricas são independentes trata as variáveis como nominais. Outros testes são geralmente mais poderosos quando as variáveis são ordinais. Esta seção apresenta um desses testes e mostra como construir intervalos de confiança para medidas ordinais de associação como o gama. As inferências se aplicam melhor a amostras aleatórias grandes. Uma regra básica é que cada C (ou D) deve exceder aproximadamente 50.

Intervalos de confiança para medidas de associação

Os intervalos de confiança ajudam a estimar a força da associação na população. Considere γ a representação do valor gama populacional. Para o gama amostral, $\hat{\gamma}$, a sua distribuição é aproximadamente normal em torno de γ. Seu erro padrão *ep* descreve a variação nos valores $\hat{\gamma}$ em torno de γ considerando amostras de um dado tamanho. A fórmula para o *ep* é complicada, mas ela é apresentada pela maioria dos *softwares*. Um intervalo de confiança para γ tem a forma:

$$\hat{\gamma} \pm z(ep).$$

EXEMPLO 8.8 Associação entre renda e felicidade

Para os dados da Tabela 8.15 sobre a renda familiar e felicidade, $\hat{\gamma} = 0,228$. Veremos na Tabela 8.18 que o $ep = 0,114$. Um intervalo de 95% de confiança para γ é então:

$$\hat{\gamma} \pm 1,96(ep), \text{ ou } 0,228 \pm 1,96(0,114), \text{ ou } 0,228 \pm 0,223,$$

que é igual a (0,005, 0,45). Podemos estar 95% confiantes de que γ não é menor do que 0,005 e nem maior do que 0,45. É plausível que essencialmente não exista associação entre renda e felicidade, mas é também plausível que exista uma associação moderada. Precisamos de um tamanho da amostra maior para estimar isso com maior precisão. ■

Teste de independência usando gama

A seguir iremos considerar um teste de independência que trata as variáveis como ordinais. Como no teste qui-quadrado, a hipótese nula é que as variáveis são estatisticamente independentes. Expressamos o teste em termos de gama, mas uma abordagem similar funciona com outras medidas ordinais de associação. A hipótese alternativa pode tomar a forma bilateral H_a: $\gamma \neq 0$ ou uma forma unilateral, H_a: $\gamma > 0$ ou H_a: $\gamma < 0$, quando prevemos a direção da associação.

A estatística-teste tem a forma da estatística z. Ela toma a diferença entre $\hat{\gamma}$ e o valor de 0 que gama assume quando H_0: independência é verdadeira e divide pelo erro padrão:

$$z = \frac{\hat{\gamma} - 0}{ep}.$$

A estatística-teste tem aproximadamente uma distribuição normal padrão quando H_0 é verdadeira. Alguns *softwares* apresentam, também, um *ep* ou um valor-*p* relacionado que vale apenas sob H_0.

EXEMPLO 8.9 Testando independência entre renda e felicidade

A Tabela 8.15, relacionando a renda familiar e a felicidade, sugere que essas variáveis estão associadas na população? O teste qui-quadrado de independência tem $\chi^2 = 3,82$ com $gl = 4$, para o qual o valor-*p* é igual a 0,43. Este teste não mostra evidência de uma associação. O teste qui-quadrado trata a variável como nominal, entretanto, e métodos de nível ordinal são mais poderosos se existe uma tendência positiva ou negativa.

A Tabela 8.18 mostra a saída para a análise da Tabela 8.15. O valor $\hat{\gamma} = 0,228$ tem $ep = 0,114$, rotulado como *Asymp. std. error* (erro padrão assintótico), onde *Asymp.* significa "assintótico" ou "amostra grande". A estatística-teste é igual a:

$$z = \frac{\hat{\gamma} - 0}{ep} = \frac{0,228 - 0}{0,114} = 2,00.$$

Da tabela normal padrão, o valor-*p* para H_a: $\gamma \neq 0$ é igual a 0,046. (O SPSS informa um valor-*p* de 0,050, com base em erro padrão diferente na estatística-teste que somente se aplica sob H_0.)

Este teste mostra evidência de uma associação. Visto que o valor amostral de gama era positivo, parece que existe uma associação positiva entre renda e felicidade. O teste para H_a: $\gamma > 0$ tem valor-p = 0,023 (ou 0,025 usando o *ep* nulo). ■

Testes ordinais *versus* teste qui-quadrado de Pearson

O resultado do teste z para estes dados fornecendo evidência de uma associação pode parecer surpreendente. A estatística qui-quadrado de $\chi^2 = 3,82$ com $gl = 4$ não forneceu evidência (valor-p = 0,43).

Um teste de independência baseado em uma medida ordinal é geralmente preferível ao teste qui-quadrado quando ambas as variáveis são ordinais. A estatística χ^2 ignora a ordenação das categorias, assumindo o mesmo valor não importando como os níveis estão ordenados. Se existe uma tendência positiva ou negativa, medidas ordinais são geralmente mais poderosas para detectá-la. Infelizmente, a situação não está bem definida. É possível para o teste qui-quadrado ser mais poderoso mesmo se os dados forem ordinais.

Para explicar isso, primeiro observamos que a hipótese nula de independência não é equivalente a um valor de 0 para o gama populacional. Embora independência implique $\gamma = 0$, o inverso não é verdadeiro. A saber, γ pode ser igual a 0 embora as variáveis

☑ **Tabela 8.18** Parte de uma saída de computador para analisar a Tabela 8.17

	Valor	GL	Sig.Assint.
Qui-quadrado de Pearson	3,816	4	0,431
	Valor	**Erro padr. Assint.**	**Sig. Aprox.**
Gama	0,2283	0,1139	0,050

não sejam estatisticamente independentes. Por exemplo, a Tabela 8.19 mostra um relacionamento entre duas variáveis que não tem uma única tendência. Além das primeiras duas colunas existe um relacionamento positivo, visto que y aumenta quando x aumenta. Além das duas últimas colunas existe um relacionamento negativo, y diminui quando x aumenta. Para toda a tabela, $C = 25(25 + 25) = 1250 = D$, assim $\gamma = 0$. A proporção de pares concordantes é igual à proporção de pares discordantes. Entretanto, não existe independência porque a distribuição condicional em y para o nível baixo de x é completamente diferente da distribuição condicional em y para o nível alto de x.

Portanto, uma medida ordinal de associação pode ser igual a 0 quando as variáveis são estatisticamente dependentes, mas a dependência não tem uma tendência geral positiva ou negativa. O teste qui-quadrado pode ter um desempenho melhor do que um teste ordinal quando o relacionamento não tem uma única tendência. Na prática, a maioria dos relacionamentos com variáveis ordinais tem geralmente uma única tendência, se existir alguma. Assim, o teste ordinal é geralmente mais poderoso do que o teste qui-quadrado.

Métodos de inferência similares para outras medidas ordinais

Os métodos de inferência para o gama também se aplicam para outras medidas ordinais de associação. Para um intervalo de confiança, tome o valor amostral e some e subtraia um escore-z vezes o erro padrão, que pode ser determinado com o uso de um *software*. Os resultados do teste são geralmente similares para qualquer medida ordinal baseada na diferença entre os números de pares concordantes e pares discordantes, como o gama e o tau-b de Kendall.

Uma abordagem alternativa para detectar tendências atribui escores às categorias para cada variável e usa a correlação e um teste z com base nela. (A Seção 9.5 apresenta um teste estreitamente relacionado.) Alguns *softwares* apresentam isso como um teste de *associação linear por linear*.

Sempre que possível, é melhor escolher as categorias para variáveis ordinais de forma precisa do que aproximada. Por exemplo, é melhor usar quatro ou cinco categorias do que somente duas. Para um determinado tamanho amostral, os erros padrão das medidas tendem a ser menores com mais categorias. Portanto, quanto melhores as categorizações, menor tende a ser o intervalo de confiança para uma medida populacional de associação. Além disso, melhores mensurações tornam mais válido o tratamento dos dados como quantitativos e o uso de métodos mais poderosos apresentados no capítulo seguinte para variáveis quantitativas.

Tabelas de contingência mistas ordinais-nominais

Para uma classificação cruzada de uma variável ordinal com uma variável nominal que tem apenas duas categorias, as medidas ordinais de associação ainda são válidas. Neste caso, o sinal da medida indica qual nível da variável nominal está associado a níveis mais altos na variável ordinal. Por exemplo, suponha gama = −0,12 para a associação em uma tabela 2×3 relacionando gênero (mulher, homem) com felicidade (pouco feliz, feliz, muito feliz). Visto que o sinal é negativo, o nível "mais alto" do gênero (isto é, homem) tende a ocorrer com o nível pouco feliz. A associação é fraca, entretanto.

Quando a variável nominal apresenta mais do que duas categorias, é inapropriado usar uma medida ordinal como o gama. Existem métodos especializados para tabelas mistas ordinais-nominais, mas é mais simples tratar a variável ordinal como quantitativa designando escores para os seus níveis. Os métodos do Capítulo 12, que generalizam comparações de duas médias para vários grupos, são, então, apropriados. A Seção 15.4 apresenta uma abordagem de modelagem que não requer a atribuição de escores às variáveis resposta ordinais.

☑ **Tabela 8.19** Um relacionamento para o qual medidas ordinais de associação se igualam a 0. As variáveis são dependentes embora o gama seja igual a 0

		Nível de y			
		Muito baixo	Baixo	Alto	Muito alto
Nível de x	Baixo	25	0	0	25
	Alto	0	25	25	0

8.7 RESUMO DO CAPÍTULO

Este capítulo introduziu análises de associação para variáveis categóricas:

- Pela *descrição das frequências* em **tabelas de contingência** utilizando distribuições de percentuais, denominadas **condicionais**, por meio das categorias da variável resposta. Se as distribuições condicionais da população são idênticas, as duas variáveis são **estatisticamente independentes** – a probabilidade de qualquer resposta particular é a mesma para cada nível da variável explicativa.
- Usando o **qui-quadrado** para **testar H_0: independência** entre as variáveis. A estatística-teste χ^2 compara cada frequência observada f_o à frequência esperada f_e satisfazendo H_0, usando:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}.$$

A estatística-teste tem uma distribuição qui-quadrado para amostras grandes. Os **graus de liberdade** dependem do número de linhas l e do número de colunas c, calculado por $gl = (l - 1)(c - 1)$. O valor-p é a probabilidade da cauda direita do valor observado χ^2.

- *Descrevendo o padrão da associação* usando **resíduos padronizados** para as células na tabela. Um resíduo padronizado apresenta o número de erros padrão que $(f_o - f_e)$ está de 0. Um valor maior do que aproximadamente 2 ou 3 em valor absoluto indica que aquela célula fornece evidência da associação em uma determinda direção.
- *Descrevendo a força da associação.* Para tabelas 2 × 2, a **diferença das proporções** é útil, como é a **razão de chances**, quociente das chances de duas linhas. Cada chance mensura a proporção de sucesso dividida pela proporção de fracassos. Quando existe independência, a diferença das proporções é igual a 0 e a razão de chances é igual a 1. Quanto mais forte a associação, mais longe estão as medidas desses valores básicos.

Este capítulo também apresentou métodos para analisar a associação entre duas variáveis ordinais.

- Muitas **medidas ordinais de associação** usam os números de **pares concordantes** (os sujeitos maiores em x são também maiores em y) e **pares discordantes** (os sujeitos maiores de x são os menores de y).
- Dos pares que são concordantes ou discordantes, o **gama** é igual à diferença entre as proporções dos dois tipos. Gama está entre −1 e +1, com valores absolutos maiores indicando uma associação mais forte. Quando as variáveis são independentes, gama é igual a 0.

O teste qui-quadrado trata os dados como nominais. Quando as variáveis são ordinais, os métodos que usam essa informação (como um teste z baseado no gama amostral) são mais poderosos para detectar uma tendência de associação positiva ou negativa.

O próximo capítulo introduz métodos similares para descrever e fazer inferências sobre a associação entre duas variáveis quantitativas.

EXERCÍCIOS

Praticando o básico

8.1 Os levantamentos de dados da PSG rotineiramente mostram que, nos Estados Unidos, aproximadamente 40% dos homens e 40% das mulheres acreditam que uma mulher deveria ser capaz de fazer um aborto se ela quiser por qualquer motivo (variável "ABANY").
- **(a)** Construa uma tabela de contingência mostrando a distribuição condicional da variável: o aborto sem restrição deveria ser legal (sim, não) por gênero.
- **(b)** Com base nesses resultados, a independência estatística parece plausível entre o gênero e a opinião sobre o aborto sem restrição? Explique.

8.2 Uma mulher ficar grávida no próximo ano é uma variável categórica com categorias (sim, não) e se ela e seu parceiro usam contraceptivos é outra variável categórica com categorias (sim, não). Você esperaria que essas variáveis fossem estatisticamente independentes ou associadas? Explique.

8.3 A cada ano, uma pesquisa em larga escala com calouros da universidade é conduzida pelo Higher Education Research Institute (Instituto de Pesquisa em Educação Superior) da UCLA que pergunta as opiniões dos calouros sobre assuntos variados. Em 2002, 46% dos homens e 35% das mulheres da pesquisa de 283000 calouros da universidade indicaram apoio para a legalização da maconha.
- **(a)** Se os resultados para a população de calouros da universidade fossem similares a estes, o gênero e a opinião sobre a legalização da maconha seriam independentes ou dependentes?
- **(b)** Determine os percentuais hipotéticos populacionais em uma tabela de contingência na qual estas variáveis fossem independentes.

8.4 Alguns analistas políticos afirmam que, durante a presidência de George W. Bush, a popularidade dos Estados Unidos diminuiu dramaticamente em todo o mundo. Em *America Against the World: How We Are Different and Why We Are Disliked*[1] (A América contra o mundo: como somos diferentes e por que não gostam de nós), o Pew Research Center (Centro de Pesquisas Pew) resumiu os resultados de 91000 entrevistas realizadas em 51 países. Na Alemanha, por exemplo, o estudo mostrou que os que tinham uma opinião favorável sobre os Estados Unidos mudaram entre 2000 e 2006 de 78% para 37%. Mostre como construir uma tabela de contingência relacionando a opinião a respeito dos Estados Unidos para os dois anos do levantamento de dados, para a Alemanha. Para esta tabela, identifique a variável resposta, a variável explicativa e as distribuições condicionais.

8.5 Baseado em estimativas atuais sobre quão bem as mamografias detectam o câncer de mama, a Tabela 8.20 mostra o que esperar para 100000 mulheres adultas acima dos 40 anos em termos de se uma mulher tem câncer de mama e um resultado positivo da mamografia (isto é, indica que uma mulher tem câncer de mama).
- **(a)** Construa as distribuições condicionais para o resultado do teste da mamografia, dado o *status* verdadeiro da doença. A mamografia parece ser uma boa ferramenta de diagnóstico?
- **(b)** Construa a distribuição condicional do *status* da doença para as que têm um resultado positivo no teste. Use isto para explicar por que mesmo um bom diagnóstico do teste pode ter uma alta taxa de falsos positivos quando a doença não é comum.

☑ **Tabela 8.20**

		Diagnóstico do teste	
		Positivo	Negativo
Câncer de mama	Sim	860	140
	Não	11800	87120

8.6 Os dados publicados no *site* do FBI (www.fbi.gov) indicaram que, de todos os negros assassinados em 2005, 91% foram assassinados por negros, e, de todos os brancos assassinados em 2005, 83% foram assassinados por brancos. Considere y a representação da raça da vítima e x a representação da raça do assassino.
- **(a)** A quais distribuições condicionais estas estatísticas se referem, as de y para os níveis dados de x ou as de x para os níveis dados de y? Construa uma tabela de contingência mostrando estas distribuições.
- **(b)** x e y são independentes ou dependentes? Explique.

8.7 Que valores do χ^2 fornecem valores-p de 0,05 para testar a independência considerando tabelas com as seguintes dimensões?

(a) 2×2 **(b)** 3×3 **(c)** 2×5
(d) 5×5 **(e)** 3×9

8.8 Verifique que a Tabela de Contingência 8.21 tem quatro graus de liberdade, mostrando como as quatro frequências fornecidas determinam as demais.

☑ **Tabela 8.21**

10	20		60
30	40		100
			40
50	80	70	

8.9 Em 2000 a PSG perguntou se uma pessoa estaria disposta a aceitar cortes no seu padrão de vida para ajudar o meio ambiente ("GRNSOL"), com categorias (muito disposto, moderadamente disposto, indiferente, pouco disposto, nada disposto). Quando estes resultados foram cruzados com o sexo da pessoa, o valor χ^2 foi de 8,0.
- **(a)** Quais são as hipóteses para o teste referido?
- **(b)** Relate o valor dos *gl* no qual o χ^2 está baseado.
- **(c)** A que conclusão você chegaria usando um nível de significância de (i) 0,05, (ii) 0,10? Formule as suas conclusões no contexto desse estudo.

8.10 A Tabela 8.22 se refere a um levantamento de dados com estudantes do último ano do ensino médio de Dayton, Ohio.
- **(a)** Construa distribuições condicionais que tratam o fumar como a variável resposta. Interprete.
- **(b)** Teste se fumar e beber são estatisticamente dependentes. Determine o valor-p e interprete.

☑ **Tabela 8.22**

		Fuma	
		Sim	Não
Bebe	Sim	1449	500
	Não	46	281

Fonte: Agradecimentos ao Professor Harry Khamis pelo fornecimento desses dados.

8.11 As pessoas que acreditam na vida após a morte são mais felizes? Vá ao *site* da PSG sda.berkeley.edu/GSS e baixe uma tabela de contingência para o levantamento de dados de 2006 relacionando felicidade e crença na vida após a morte (variáveis "HAPPY" e "POSTLIFE", com "YEAR(2006)" em [*selection filter*]).
- **(a)** Formule uma questão de pesquisa que poderia ser resolvida com a saída.
- **(b)** Determine as distribuições condicionais usando felicidade como a variável resposta e interprete.
- **(c)** Determine o valor do χ^2 e seu valor-p. (Você pode obter esses valores verificando em [*Statistics*].) Interprete.
- **(d)** Interprete os resíduos padronizados. (Você pode obtê-los verificando em [*z-statistic*].)

8.12 Na PSG, foi perguntado aos sujeitos, que eram casados, sobre felicidade no

casamento, a variável foi codificada como "HAPMAR".

(a) Vá a sda.berkeley.edu/GSS/ e construa uma tabela de contingência, para 2006, relacionando "HAPMAR" com a renda familiar mensurada como (acima da média, na média e abaixo da média), entrando com "FINRELA" (r: 1-2; 3; 4-5) como a variável linha e "YEAR(2006)" no filtro de seleção. Use uma tabela ou gráfico com distribuições condicionais para descrever a associação.

(b) Verificando [*Statistics*], você solicita a estatística qui-quadrado. Anote o valor bem como o do *gl* e do valor-*p* e interprete.

8.13 A amostra na Tabela 8.15 é de 157 norte-americanos negros. A Tabela 8.23 mostra as frequências e os resíduos padronizados para a renda familiar e a felicidade para sujeitos brancos na PSG de 2004.

☑ **Tabela 8.23**

Linhas: renda familiar — Colunas: felicidade

	Pouco feliz	Feliz	Muito feliz	Total
Abaixo da média	62	187	45	294
	5,34	3,43	-7,40	
Na média	47	270	181	498
	-2,73	-0,57	2,53	
Acima da média	22	127	116	267
	-2,37	-2,88	4,73	
Total	131	584	131	1059

Conteúdos das Células: Frequências
Resíduos padronizados

Qui-quadrado de Pearson = 72,15, *gl* = 4, valor-p = 0,000

(a) Explique como interpretar a estatística qui-quadrado de Pearson e seu valor-*p* associado.

(b) Explique como interpretar os resíduos padronizados das quatro células dos cantos.

8.14 A Tabela 8.24 mostra a análise do SPSS com a PSG de 2004, para variáveis identificação partidária e raça.

(a) Relate a frequência esperada para a primeira célula e mostre como o SPSS a obteve.

(b) Teste a hipótese de independência entre identificação partidária e raça. Determine a estatística-teste, o valor-*p* e interprete.

(c) Use os resíduos padronizados (aqui rotulados de ADJRES para os "resíduos ajustados" para descrever o padrão da associação).

☑ **Tabela 8.24**

Frequência PARTYID
Valor Esperado
Resíduo Ajustado

		democr	indep	republ	Total da linha
RACE	negra	250	106	17	373
		129,1	129,0	114,9	
		14,2	-2,7	-11,9	
	branca	640	783	1775	2198
		760,9	760,0	677,1	
		-14,2	2,7	11,9	
Total da coluna		890	889	792	2571

Qui-quadrado	Valor	GL	Significância
Pearson	234,73	2	0,0000

8.15 Para uma classificação cruzada 2 × 4 entre gênero e religiosidade (muito, moderadamente, levemente, de modo algum) para dados recentes da PSG, o resíduo padronizado foi de 3,2 para mulheres que são muito religiosas, −3,2 para homens que são muito religiosos, −3,5 para mulheres que não são de modo algum religiosas e 3,5 para homens que não são de modo algum religiosos. Todos os outros resíduos padronizados estavam entre −1,1 e 1,1. Interprete.

8.16 A Tabela 8.25 é da Pesquisa Social Geral de 2006, tendo uma classificação cruzada entre felicidade ("HAPPY") e estado civil ("MARITAL").

☑ **Tabela 8.25**

Estado civil	Muito feliz	Feliz	Pouco feliz
Casado	600 (13,1)	720 (−5,4)	93 (−10,0)
Viúvo	63 (−2,2)	142 (−0,2)	51 (3,4)
Divorciado	93 (−6,1)	304 (3,2)	88 (3,6)
Separado	19 (−2,7)	51 (21,2)	31 (5,3)
Nunca casou	144 (−7,4)	459 (4,2)	127 (4,0)

(a) O *software* determinou que $\chi^2 = 236{,}4$. Interprete.

(b) A Tabela 8.25 também mostra, em parênteses, os resíduos padronizados. Resuma a associação indicando qual estado civil tem uma forte evidência de que (i) mais, (ii) menos pessoas na população estejam na categoria *muito feliz* do que se as variáveis fossem independentes.

(c) Compare os grupos casado e divorciado pela diferença das proporções na categoria *muito feliz*.

8.17 Em uma pesquisa do *USA Today/Gallup* de julho de 2006, 82% dos Republicanos aprovavam o desempenho do presidente George W. Bush, enquanto 9% dos Democratas aprovavam. Você caracterizaria a associação entre a afiliação partidária e opinião sobre Bush como fraca ou forte? Explique por quê.

8.18 Em uma PSG recente, a pena de morte para sujeitos condenados por assassinato era aprovada por 74% dos brancos e 43% dos negros. Ela era aprovada por 75% dos homens e por 63% das mulheres. Nessa amostra, que variável está mais fortemente associada com a opinião sobre a pena de morte, a raça ou o gênero? Explique por quê.

8.19 Considere o Exercício 8.10, sobre o o hábito de fumar e beber.

(a) Descreva a força da associação usando a diferença entre usuários e não usuários de álcool nas proporções daqueles que fumam. Interprete.

(b) Descreva a força da associação usando a diferença entre fumantes e não fumantes nas proporções dos que bebem. Interprete.

(c) Descreva a força da associação usando a razão de chances. Interprete. O valor da razão de chances depende da sua escolha da variável resposta?

8.20 A Tabela 8.26 faz uma classificação cruzada de 68694 passageiros de carros e pequenos caminhões envolvidos em acidentes, no estado do Maine, verificando se eles estavam usando cinto de segurança e se eles tiveram ferimentos ou morreram. Descreva a associação usando:

(a) A diferença entre duas proporções tratando ferido ou morto como a variável resposta.

(b) A razão de chances.

☑ **Tabela 8.26**

		Ferido ou morto	
		Sim	Não
Cinto de Segurança	Sim	2409	35383
	Não	3865	27037

Fonte: Agradecimentos a Dra. Cristanna Cook, Medical Care Development, Augusta, Maine, pelo fornecimento destes dados.

8.21 De acordo com o Substance Abuse and Mental Health Archive (Arquivo de Abuso de Substâncias e Saúde Mental), um levantamento nacional por domicílio, em 2003, sobre o uso de drogas, indicou que para norte-americanos com idades entre 26-34, 51% tinham usado maconha pelo menos uma vez na vida e 18% tinham usado pelo menos uma vez cocaína.

(a) Encontre a chances de alguém ter usado (i) maconha, (ii) cocaína. Interprete.

(b) Encontre a razão de chances comparando o uso da maconha ao da cocaína. Interprete.

8.22 De acordo com o U.S. Department of Justice (Departamento de Justiça dos Estados Unidos), em 2004, a taxa de detentos, nas prisões do país, era de 1 por 109 detentos homens, 1 por 1563 detentas mulheres, 1694 por 100000 detentos negros e 252 por 100000 detentos brancos (*Fonte*: www.ojp.usdoj.gov/bjs).

(a) Encontre a razão de chances entre detento e (i) gênero, (ii) raça. Interprete.

(b) De acordo com a razão de chances, qual tem a associação mais forte com estar preso, gênero ou raça? Explique.

8.23 Considere a Tabela 8.1 (página 253) sobre identificação partidária e gênero. Encontre e interprete a razão de chances para cada subtabela 2×2. Explique por que esta análise sugere que as duas últimas colunas não mostram essencialmente associação.

8.24 Para calouros universitários, em 2004, o percentual que concordava que relacionamentos homossexuais deveriam ser legalmente proibidos era de 38,0% dos homens e 23,4% dos mulheres (www.gseis.ucla.edu/heri/american_freshman.html).

(a) A razão de chances é de 2,01. Explique o que está errado com a interpretação: "A probabilidade de uma resposta *sim* para homens é 2,01 vezes a probabilidade de uma resposta *sim* para mulheres". Dê a interpretação correta.

(b) A chance de uma resposta *sim*, para homens, é igual a 0,613. Estime a probabilidade de uma resposta *sim* para os homens.

(c) Baseado na chance de 0,613 para homens e a razão de chances de 2,01, mostre como estimar a probabilidade de uma resposta *sim* para as mulheres.

8.25 A Tabela 8.27 faz uma classificação cruzada entre felicidade e renda familiar para a subamostra da PSG de 2004 que se identificou como judeu.

☑ **Tabela 8.27**

		FELIZ		
		Pouco	Moderadamente	Muito
RENDA	Abaixo da média	1	2	1
	Na média	0	5	2
	Acima da média	2	4	0

(a) Encontre o número de (i) pares concordantes, (ii) pares discordantes.

(b) Encontre gama e interprete.

(c) Mostre como expressar gama como a diferença entre duas proporções.

8.26 Para a PSG de 2006, $\hat{\gamma} = 0{,}22$ para o relacionamento entre satisfação no emprego ("SATJOB" – categorizada como muito insatisfeito, insatisfeito, moderadamente satisfeito e muito satisfeito) e renda familiar ("FINRELA" – categorizada como abaixo da média, na média e acima da média).

(a) Você consideraria isto uma associação muito forte ou relativamente fraca? Explique.

(b) Dos pares que são concordantes ou discordantes, que proporção é concordante? Discordante?

(c) Esta é uma associação mais forte ou mais fraca do que aquela entre a satisfação no emprego e felicidade (variável "HAPPY"), que tem $\hat{\gamma} = 0{,}40$? Explique.

8.27 Um estudo sobre aspirações educacionais de estudantes do ensino médio[2] mensurou as aspirações (usando a escala: ensino médio incompleto, ensino médio completo, graduação incompleta, graduação completa) e renda familiar com três categorias ordenadas. O *software* forneceu os resultados mostrados na Tabela 8.28.

(a) Use o gama para resumir a associação.

(b) Verifique a independência das aspirações educacionais e da renda familiar usando o teste qui-quadrado. Interprete.

(c) Encontre o intervalo de 95% para gama. Interprete.

(d) Conduza um teste alternativo de independência que leve em consideração a ordenação das categorias. Por que os resultados são tão diferentes do teste qui-quadrado?

☑ **Tabela 8.28**

Estatística	GL	Valor	Prob
Qui-quadrado	6	8,871	0,181
Estatística		**Valor**	**ASE***
Gama		0,163	0,080

8.28 Considere o Exercício 8.13 sobre a relação entre felicidade e renda. A análise feita lá não levou em consideração a ordinalidade das variáveis. Usando *software*:

(a) Resuma a força da associação encontrando e interpretando o gama.

(b) Construa e interprete um intervalo de 95% de confiança para o valor populacional do gama.

Conceitos e aplicações

8.29 Considere o arquivo do levantamento de dados com os estudantes (Exercício 1.11, página 25). Usando um *software*, crie e analise de forma descritiva a tabela de contingência relacionando a opinião sobre o aborto e **(a)** afiliação política, **(b)** religiosidade.

8.30 Considere o arquivo de dados que você criou no Exercício 1.12 (página 26). Para as variáveis escolhidas pelo seu professor, proponha uma pergunta de pesquisa e conduza análises estatísticas inferenciais e descritivas. Interprete e resuma o que você descobriu em um breve relatório.

8.31 Em 2002, a PSG perguntou como o trabalho doméstico era divido entre o respondente e o seu ou sua parceiro(a) ("HHWKFAIR"). As respostas possíveis foram 1 = eu faço muito mais do que a minha cota justa, 2 = eu faço um pouco mais do que a minha cota justa, 3 = eu mal faço a minha cota justa, 4 = eu faço um pouco menos do que a minha cota justa, 5 = eu faço bem menos do que a minha cota justa. A Tabela 8.29 mostra os resultados de acordo com o sexo do respondente. Proponha uma pergunta de pesquisa que poderia ser feita com esta saída e prepare um relatório de uma página resumindo o que você descobriu. (O "Resíduo ajust". é o resíduo padronizado.)

8.32 Proponha uma pergunta de pesquisa sobre a atitude em relação a relações homossexuais e ideologia política. Usando os dados mais recentes da PSG em "HOMOSEX" e "POLVIEWS", conduza uma análise descritiva e inferencial para tratar dessa questão. Prepare um breve relatório resumindo suas análises.

8.33 Vários sociólogos têm relatado que o preconceito racial varia de acordo com o grupo religioso. Examine isto usando a Tabela 8.30, para respondentes brancos da PSG de 2002. As variáveis são Fundamentalismo *versus* Liberalismo da Religião do Respondente ("FUND") e a reposta à pergunta ("RACMAR"): "Você acha que deveria haver leis contra casamentos entre negros e brancos?". Analise esses dados. Prepare um relatório, descrevendo as suas análises e interprete os dados.

* N. de T. T.: *Assymptotic Standard Error* ou Erro Padrão Assintótico.

☑ **Tabela 8.29**

			HHWFAIR					Total
			1	2	3	4	5	
sexo	feminino	frequência	121	108	135	19	6	389
		% entre sexo	31,1%	27,8%	34,7%	4,9%	1,5%	100,0%
		Resíduo ajust.	8,0	5,9	-4,2	-7,1	-4,9	
	masculino	frequência	18	28	148	68	29	291
		% entre sexo	6,2%	9,6%	50,9%	23,4%	10,0%	100,0%
		Resíduo ajust.	-8,0	-5,9	4,2	7,1	4,9	

	Valor	gl	Sig. Assint
Qui-quadrado de Pearson	155,8	4	0,000
	Valor	**Erro padrão Assint.**	
Gama	0,690	0,038	

☑ **Tabela 8.30**

Preferência religiosa	Leis contra o casamento		
	A favor	Contra	Total
Fundamentalista	39	142	181
Moderado	21	248	269
Liberal	17	236	253
Nenhuma	16	74	90
Total	93	700	793

8.34 Para os dados de 2006 da PSG, daqueles que se identificaram como Democratas, 616 se classificaram como liberais e 262 como conservadores. Daqueles que se identificaram como Republicanos, 94 se denominaram liberais e 721 conservadores. Usando métodos apresentados nesse capítulo, descreva a força da associação.

8.35 Um estudo[3] das forças armadas norte-americanas que serviram no Iraque ou Afeganistão descobriu que o evento de ser atacado ou emboscado foi relatado por 1139 de 1961 membros do Exército que serviram no Afeganistão, 789 de 883 membros do Exército que serviram no Iraque e 764 de 805 fuzileiros navais que serviram no Iraque. Resuma esses dados usando distribuições condicionais e medidas de associação.

8.36 Um pouco antes de uma eleição governamental, uma pesquisa fez as seguintes perguntas a uma amostra aleatória de 50 eleitores em potencial:

Você se considera um Democrata (D), um Republicano (R) ou um Independente (I)?

Se você fosse votar hoje, você votaria no candidato Democrata (D), no candidato Republicano (R) ou você estaria indeciso (I) sobre como votar?

Você planeja votar na eleição? Sim (S) ou Não (N)?

Para cada pessoa entrevistada, as respostas às três perguntas são colocadas em um arquivo de dados. Por exemplo, o registro (D, I, N) representa um Democrata registrado que está indeciso e não espera votar. A Tabela 8.31 resume os resultados dos 50 entrevistados. Usando um *software*, crie um arquivo de dados e execute as seguintes análises:

(a) Construa a tabela de contingência de 3 × 3 relacionando a afiliação partidária com intenção de voto. Relate as distribuições condicionais das intenções de voto para cada uma das três afiliações partidárias. Elas são muito diferentes?

(b) Relate o resultado do teste de hipóteses em que a intenção de voto é independente da afiliação partidária. Forneça a estatística-teste, o valor-p e interprete o resultado.

(c) Complemente a análise em (a) – (b) para investigar melhor a associação. Interprete.

☑ **Tabela 8.31**

(D, I, N)	(R, R, S)	(I, D, S)	(I, I, N)	(R, I, N)
(I, D, N)	(R, R, S)	(I, I, N)	(D, I, S)	(D, R, N)
(I, D, N)	(D, D, S)	(D, D, S)	(I, D, S)	(R, I, N)
(D, R, N)	(R, D, N)	(D, I, N)	(D, D, S)	(R, R, S)
(R, R, S)	(D, D, N)	(D, D, S)	(I, D, S)	(R, R, N)
(D, D, S)	(D, R, S)	(I, I, N)	(D, D, N)	(D, D, S)
(R, R, S)	(R, R, S)	(D, I, N)	(I, R, N)	(I, R, S)
(R, R, S)	(I, I, S)	(D, D, S)	(D, R, S)	(D, D, N)
(D, D, S)	(I, R, S)	(R, R, S)	(I, D, S)	(R, R, N)
(R, R, S)	(D, D, S)	(I, D, S)	(I, R, N)	(R, R, S)

8.37 **(a)** Quando o tamanho da amostra é muito grande, não determinamos necessariamente um resultado importante quando mostramos uma associação estatisticamente significativa. Explique.

(b) As observações nas Seções 8.3 e 8.4 sobre valores-p pequenos não querendo dizer necessariamente um efeito importante se aplicam a qualquer teste de significância. Explique por que, discutindo o efeito do valor de n nos erros padrão e nos tamanhos das estatísticas-teste.

8.38 Responda verdadeiro ou falso ao que segue. Explique sua resposta.

(a) Mesmo quando as distribuições condicionais da amostra em uma tabela de contingência são somente um pouco diferentes, quando o tamanho da amostra é muito grande é possível ter uma estatística-teste χ^2 grande e um valor-p pequeno para testar H_0: independência.

(b) Se a razão de chances = 2,0 entre gênero (homem, mulher) e opinião em algum assunto (a favor, contra), então a razão de chances = −2,0 se mensurarmos gênero como (mulher, homem).

(c) Trocar duas linhas em uma tabela de contingência não tem efeito na estatística qui-quadrado.

(d) Trocar duas linhas em uma tabela de contingência não tem efeito no gama.

(e) Se $\gamma = 0$ entre duas variáveis, então as variáveis são estatisticamente independentes.

8.39 A resposta correta no Exercício 8.38 (c) implica que se a estatística qui-quadrado é usada para uma tabela de contingência tendo categorias ordenadas em ambas as direções, então (selecione a(s) opção(ões) correta(s)):

(a) A estatística realmente trata as variáveis como nominais.

(b) A informação sobre a ordem é ignorada.

(c) O teste não é geralmente tão poderoso para detectar associação quanto uma estatística-teste baseada nos números de pares concordantes e discordantes.

(d) A estatística não pode diferenciar entre associações positivas e negativas.

8.40 Cada sujeito em uma amostra de 100 homens e 100 mulheres é solicitado a indicar quais dos seguintes fatores (um ou mais) são responsáveis pelos aumentos em crimes cometidos por adolescentes: A – a crescente distância na renda entre ricos e pobres, B – o aumento de famílias de pai ou mãe solteiro(a), C – tempo insuficiente que os pais passam com seus filhos, D – as penalidades criminais dadas pela corte judicial são muito tolerantes, E – os problemas crescentes com drogas na sociedade, F – níveis crescentes de violência mostrados na televisão. Para analisar se as respostas diferem por gênero do respondente, foi feita uma classificação cruzada das respostas por gênero, como a Tabela 8.32 mostra.

(a) É válido aplicar o teste qui-quadrado de independência para estes dados? Explique.

(b) Explique como esta tabela realmente fornece a informação necessária para fazer a classificação cruzada de gênero com cada uma das seis variáveis. Construa a tabela de contingência relacionando gênero à opinião sobre se a crescente distância na renda é responsável pelo crescimento dos crimes cometidos por adolescentes.

☑ **Tabela 8.32**

Gênero	A	B	C	D	E	F
Homens	60	81	75	63	86	62
Mulheres	75	87	86	46	82	83

***8.41** A Tabela 8.33 exibe a associação máxima possível entre duas variáveis binárias para uma amostra de tamanho n.

(a) Mostre que $\chi^2 = n$ para essa tabela e, portanto, que o valor máximo de χ^2 para tabelas 2 × 2 é n.

(b) A medida de associação **phi-quadrado** para tabelas de contingência 2 × 2 tem um valor amostral de :

$$\hat{\phi}^2 = \frac{\chi^2}{n}.$$

Explique por que esta medida está entre 0 e 1, com um valor populacional igual a 0 correspondendo à independência. (É um caso especial, para tabelas 2 × 2, da medida *tau de Goodman e Kruskal* e a medida r^2 a ser introduzida no próximo capítulo.)

☑ **Tabela 8.33**

$n/2$	0
0	$n/2$

***8.42** Para tabelas 2 × 2, o coeficiente gama pode ser simplificado para uma medida proposta inicialmente em torno de 1900 pelo estatístico G. Udny Yule, que também introduziu a razão de chances. Neste caso especial, gama é chamado de **Q de Yule**.

(a) Mostre que para uma tabela genérica com frequências (a, b) na linha 1 e (c, d) na linha 2, o número de pares concordantes é igual a ad, o número de pares discordantes é igual a bc e $Q = (ad - bc)/(ad + bc)$.

(b) Mostre que o valor absoluto de gama é igual a 1 para qualquer tabela 2×2 na qual uma das frequências da célula é 0.

***8.43** Construa uma tabela 3×3 para cada uma das seguintes condições:

(a) O gama é igual a 1. (*Dica*: não deve haver pares discordantes.)

(b) O gama é igual a -1.

(c) O gama é igual a 0.

***8.44** Uma variável qui-quadrado com graus de liberdade iguais a gl tem a representação $z_1^2 + \ldots + z_{gl}^2$, onde $z_1, \ldots, z_{gl}$ são variáveis normais padrão independentes.

(a) Se z é uma estatística-teste que tem uma distribuição normal padrão, que distribuição tem z^2?

(b) Explique como obter os valores do qui-quadrado para $gl = 1$ na Tabela C, dos escores-z, da tabela normal padrão (Tabela A). Ilustre com o valor do qui-quadrado de 6,63 que tem um valor-p de 0,01.

(c) A estatística qui-quadrado para testar H_0: independência entre a crença na vida após a morte (sim, não) e felicidade (pouco feliz, feliz, muito feliz) é χ_1^2 em uma tabela 2×3 para homens e χ_2^2 em uma tabela 2×3 para mulheres. Se H_0 é verdadeiro para cada gênero, então qual é a distribuição de probabilidade de $\chi_1^2 + \chi_2^2$?

***8.45** Para uma tabela 2×2 com frequências a, b, c, d, o logaritmo da razão de chances da amostra $\log\hat{\theta}$ têm uma distribuição amostral aproximadamente normal com erro padrão estimado igual a:

$$ep = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

Os anti-logaritmos dos extremos do intervalo de confiança para $\log(\theta)$ são os extremos do intervalo de confiança para θ. Para a Tabela 8.13 da página 267, mostre que $ep = 0{,}0833$ e um intervalo de 95% de confiança para a razão de chances é (67,3; 93,2). Interprete.

☑ NOTAS

[1] KOHUT, A., STOKES, B. *America Against the World: How We Are Different and Why We Are Disliked*, Times Books, 2006.

[2] CRYSDALE, S. *Intern. J. Compar. Social*, v. 16, p. 19-36, 1975.

[3] HOGE, C. et al. *New England J. Medic*, v.. 351, p. 13-21, 2004.

☑ 9 REGRESSÃO LINEAR E CORRELAÇÃO

O Capítulo 8 apresentou métodos para analisar a associação entre variáveis categóricas resposta e explicativa. Este capítulo apresenta métodos para analisar variáveis quantitativas resposta e explicativa.

A Tabela 9.1 mostra dados do *Statistical Abstract of the United States* (Resumo Estatístico dos Estados Unidos) para os 50 estados e o Distrito de Columbia (D.C.) no que segue:

- Taxa de assassinato: o número de assassinatos por 100000 habitantes.
- Taxa de crimes violentos: o número de assassinatos, estupros violentos, assaltos e agressão com circunstâncias agravantes por 100000 habitantes.
- Percentual da população com renda abaixo do nível de pobreza
- Percentual de famílias chefiadas por um único progenitor.

Para essas variáveis quantitativas, a taxa de crimes violentos e a taxa de assassinatos são variáveis respostas naturais. Trataremos a taxa de pobreza e o percentual de famílias com um único progenitor como variáveis explicativas para estas respostas à medida que estudamos os métodos para analisar os relacionamentos entre variáveis quantitativas neste capítulo e em alguns exercícios. O *site* do livro contém dois conjuntos de dados sobre estas e outras variáveis que iremos analisar em exercícios neste e em capítulos posteriores.

Analisamos três aspectos diferentes, porém relacionados, de tais relacionamentos:

1. Investigamos *se existe uma associação* entre as variáveis, testando a hipótese de independência estatística.
2. Estudamos a *força de sua associação* usando a medida de *correlação* da associação.
3. Estimamos a *equação da regressão* que prevê o valor da variável resposta a partir do valor da variável explicativa. Por exemplo, tal equação prevê a taxa de assassinatos do estado usando o percentual da população que está vivendo abaixo do nível de pobreza.

As análises são coletivamente chamadas de **análises de regressão**. A Seção 9.1 mostra como usar uma linha reta para a equação de regressão, e a Seção 9.2 mostra como usar os dados para estimar essa linha. A Seção 9.3 introduz o *modelo de regressão linear*, que leva em consideração a variabilidade dos dados em torno da linha de regressão. A Seção 9.4 usa a *correlação* e o seu quadrado para descrever a força da associação. A Seção 9.5 apresenta a inferência estatística para uma análise de regressão. A seção final faz um exame minucioso das associações e os riscos potenciais no uso da regressão.