Universidade Federal Fluminense

Instituto de Matemática e Estatística

Departamento de Estatística

Análise de Dados Categóricos

Ana Maria Lima de Farias

Fábio Nogueira Demarqui

Departamento de Estatística

Março 2017

Sumário

Capítulo 1

Análise de dados categóricos

1.1 Introdução

Vamos, agora, estudar algumas técnicas de inferências para dados qualitativos. Assim como no estudo anterior de populações normais, vamos considerar os casos de uma população, em que cada indivíduo é classificado segundo alguma variável qualitativa, e duas populações, das quais amostras retiradas podem ser independentes ou dependentes (dados emparelhados) e para cada elemento da amostra são observadas duas variáveis qualitativas. Os objetivos da inferência variam para as diferentes situações. A título de ilustração, vamos considerar os seguintes exemplos.

EXEMPLO 1.1 Meio de transporte

Em uma pesquisa com alunos da UFF em Niterói, perguntou-se a cada um deles o meio de transporte que utilizavam no trajeto de casa para a universidade. Cada aluno escolhia entre uma das seguintes possibilidades: só ônibus, só barca, ônibus e barca, carro, caminhada/bicicleta. As proporções amostrais $\widehat{P}_1$, $\widehat{P}_2$, $\widehat{P}_3$, $\widehat{P}_4$ e $\widehat{P}_5$ são estimadores das verdadeiras proporções populacionais p_1, p_2, p_3, p_4 e p_5 de usuários de cada meio de transporte e um interesse poderia ser testar se $p_1 = 0,60$, $p_2 = 0,10$, $p_3 = 0,20$, $p_4 = 0,05$ e $p_5 = 0,05$. Esse é um *teste de aderência*, ou seja, estamos testando se nossos dados são compatíveis com (aderem a) determinada distribuição.

EXEMPLO 1.2 Meio de transporte e gênero

Ainda no estudo sobre meio de transporte, um interesse poderia ser estudar se há diferença entre homens e mulheres no meio de transporte utilizado. Para isso, amostras *independentes* de homens e mulheres seriam retiradas da população de todos os estudantes da UFF em Niterói e para os indivíduos de cada amostra seria registrada a variável meio de transporte. Note que a segunda variável – gênero – definiu as populações. O interesse aqui seria testar se homens e mulheres usam igualmente os diferentes meios de transporte, ou seja, queremos testar $p_{1H} = p_{1M}$, $p_{2H} = p_{2M}$, $p_{3H} = p_{3M}$, $p_{4H} = p_{4M}$ e $p_{5H} = p_{5M}$, em que p_{iH} e p_{iM} representam as proporções de homens e mulheres que utilizam o meio de transporte i. Este é um exemplo de um *teste de homogeneidade*, ou seja, estamos testando se as populações de homens e mulheres são homogêneas em relação à variável meio de transporte. Note que em cada população a variável de interesse – meio de transporte – segue uma distribuição multinomial. Sendo assim, nosso interesse é testar se as duas distribuições multinomiais são iguais (ou homogêneas).

EXEMPLO 1.3 Meio de transporte e uso do bandejão

Continuando com o estudo sobre alunos da UFF em Niterói, outro interesse poderia ser a relação entre meio de transporte utilizado e uso do bandejão da UFF. Assim, para cada aluno da amostra seriam registradas duas variáveis: meio de transporte utilizado e uso do bandejão (Sim ou Não). Agora temos dados emparelhados (amostras dependentes) e nosso interesse é um *teste de independência* entre as variáveis.

Os três testes citados acima se basearão numa estatística que compara as frequências *observadas* O com as frequências *esperadas* E sob a suposição de veracidade da hipótese nula. Grandes diferenças entre tais frequências são indicativo de que a hipótese nula não é verdadeira. Mas para termos uma medida de distância, consideraremos as diferenças ao quadrado, isto é, $(O-E)^2$. Além disso, iremos considerar distâncias *relativas*: uma distância $(O-E)^2 = 50$ será mais relevante para um valor esperado de 30 do que para um valor esperado de 100. Dessa forma, nossa estatística de teste terá a forma final dada por

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \tag{1.1}$$

e essa estatística terá (aproximadamente) uma distribuição qui-quadrado com número de graus de liberdade que varia de acordo com o teste considerado. Note que a base de comparação é o valor esperado sob H_0, um valor fixo, bem determinado, que não depende da amostra sorteada.

Vamos, agora, detalhar cada um dos testes.

1.2 Dados univariados: Teste de aderência

Consideremos uma população descrita por uma variável categórica X que assume k valores (no Exemplo 1.1, X é o meio de transporte com $k = 5$ categorias). Sejam $p_1, p_2, \cdots, p_k$ as proporções populacionais das categorias $1, 2, \cdots, k$, respectivamente. Queremos testar

$$H_0 : p_1 = p_{10}\,,\ p_2 = p_{20}\,,\ \cdots\,,\ p_k = p_{k0} \tag{1.2}$$

$$H_1 : p_i \neq p_{i0} \text{ para pelo menos um } i \tag{1.3}$$

em que $p_{i0}, i = 1, 2, \cdots, k$ são as proporções hipotéticas que devem satisfazer $\sum_{i=1}^{k} p_{i0} = 1$.

Suponha que uma amostra de tamanho n seja selecionada de tal população. Seja N_i o número de observações da amostra pertencentes à categoria i, $i = 1, 2, \cdots, k$ (esses números mudam ao longo de todas as possíveis amostras de tamanho n). Se H_0 é verdadeira, o número esperado de observações na categoria i é

$$e_i = np_{i0} \tag{1.4}$$

Na Tabela 1.1 temos o resumo dessa situação.

Tabela 1.1 – Teste de aderência - esquema dos dados

Categoria	1	2	$\cdots$	k	Soma
Proporção populacional	p_1	p_2	$\cdots$	p_k	1
Proporção populacional sob H_0	p_{10}	p_{20}	$\cdots$	p_{k0}	1
Número observado	N_1	N_2	$\cdots$	N_k	n
Número esperado sob H_0	np_{10}	np_{20}	$\cdots$	np_{k0}	n

A estatística de teste é

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - np_{i0})^2}{np_{i0}} \underbrace{\approx}_{\text{sob } H_0} \chi^2_{k-1} \tag{1.5}$$

Tal aproximação é boa se $e_i = np_{i0} \geq 5$ para todo $i = 1, 2, \ldots, k$. Valores grandes indicam grandes afastamentos entre os valores observados e esperados; assim, a região crítica para um nível de significância α é

$$X^2 > \chi^2_{k-1;\alpha} \tag{1.6}$$

É interessante observar as seguintes propriedades sobre os valores observados e esperados.

- $\sum_{i=1}^{k} e_i = \sum_{i=1}^{k} N_i = n$ (soma dos valores esperados)

  De fato: $\sum_{i=1}^{k} e_i = \sum_{i=1}^{k} np_{i0} = n \sum_{i=1}^{k} p_{i0} = n \cdot 1 = n$

- $\sum_{i=1}^{k} (N_i - e_i) = 0$ (desvios em relação aos esperados)

  De fato: $\sum_{i=1}^{k} (N_i - e_i) = \sum_{i=1}^{k} N_i - \sum_{i=1}^{k} e_i = n - n = 0$

EXEMPLO 1.4 Meio de transporte – continuação

Uma amostra de 200 estudantes da UFF de Niterói mostrou que 65 usam apenas ônibus como meio de transporte, 7 usam apenas barca, 20 usam ônibus e barca, 5 usam carro e 2 caminham ou vão de bicicleta até a universidade. Vamos testar

$$H_0 : p_1 = 0,60\,,\ p_2 = 0,10\,, p_3 = 0,20\,,\ p_4 = p_5 = 0,05$$

Na tabela a seguir ilustram-se os cálculos necessários.

Categoria i	Valor observado O_i	Valor esperado E_i	Parcela de X^2
1=Só ônibus	124	$200 \cdot 0,6 = 120$	$\frac{(124-120)^2}{120}$
2=Só barca	22	$200 \cdot 0,1 = 20$	$\frac{(22-20)^2}{20}$
3=Ônibus e barca	35	$200 \cdot 0,2 = 40$	$\frac{(35-40)^2}{40}$
4=Carro	9	$200 \cdot 0,05 = 10$	$\frac{(9-10)^2}{10}$
5=Caminhada/bicicleta	10	$200 \cdot 0,05 = 10$	$\frac{(10-10)^2}{10}$

Todos os valores esperados são maiores que 5 e, assim, podemos usar a aproximação qui-quadrado com 4 graus de liberdade. A região crítica é $X^2 > 9,4877$ para um nível de significância $\alpha = 0,05$. O valor observado da estatística de teste é

$$x_0^2 = \frac{16}{120} + \frac{4}{20} + \frac{25}{40} + \frac{1}{10} + \frac{0}{10} = \frac{16+24+75+12+0}{120} = \frac{129}{120} = 1,075$$

Como $1,075 < 9,4877$, não rejeitamos H_0, ou seja, não há evidências de que as proporções populacionais sejam diferentes das hipotéticas.

♦♦

EXEMPLO 1.5 Duas categorias

No estudo da inferência para uma população $X \sim Bern(p_1)$, vimos que a estatística de teste para $H_0 : p_1 = p_{10}$ é

$$Z = \frac{\widehat{P} - p_0}{\sqrt{p_{10}(1-p_{10})}} \underbrace{\approx}_{\text{sob } H_0} N(0;1)$$

Como são apenas duas categorias, basta trabalhar com uma delas e a outra estará determinada. Se denotarmos por N_i o número de observações na categoria i ($i = 1$ - sucessos; $i = 2$ - fracassos) e por p_{i0} a proporção de observações na categoria i para sermos coerentes com a notação anterior em que $k \geq 2$, a aproximação normal para a binomial nos dá que

$$Z = \frac{N_1 - np_{10}}{\sqrt{np_{10}(1-p_{10})}} \underbrace{\approx}_{\text{sob } H_0} N(0;1)$$

Logo,

$$Q = Z^2 = \frac{(N_1 - np_{10})^2}{np_{10}(1-p_{10})} \underbrace{\approx}_{\text{sob } H_0} \chi_1^2$$

Notando que $(1 - p_{10}) + p_{10} = 1$, podemos reescrever Q como

$$\begin{aligned} Q = Z^2 &= \frac{(N_1 - np_{10})^2}{np_{10}(1-p_{10})}[(1-p_{10}) + p_{10}] \\ &= \frac{(N_1 - np_{10})^2}{np_{10}} + \frac{(N_1 - np_{10})^2}{n(1-p_{10})} \\ &= \frac{(N_1 - np_{10})^2}{np_{10}} + \frac{[(n - N_2) - n(1-p_{20}]^2}{np_{20})} \\ &= \frac{(N_1 - np_{10})^2}{np_{10}} + \frac{[-N_2 + np_{20}]^2}{np_{20})} \\ &= \frac{(N_1 - np_{10})^2}{np_{10}} + \frac{[N_2 - np_{20}]^2}{np_{20})} \end{aligned}$$

Vemos, assim, que há uma equivalência entre as estatísticas Z e χ^2 quando $k = 2$.

♦♦

1.3 Dados bivariados

Consideremos um exemplo em que as variáveis categóricas $X =$ "opinião sobre a legalização do uso medicinal da maconha" com 3 níveis (a favor, indiferente e contra) e $Y =$ "faixa etária " com 4 níveis (16 a

18, 19 a 25, 26 a 40 e mais de 40) serão estudadas em uma pesquisa por amostragem entre os alunos de uma grande universidade. O que veremos agora é que a forma de se coletarem os dados é fundamental para se definir o tipo de análise pertinente.

Suponhamos, inicialmente, que sejam selecionadas amostras independentes de tamanhos $n_1, \cdots, n_4$ das quatro faixas etárias e cada elemento de cada uma das 4 amostras indique sua opinião sobre a legalização do uso medicinal da maconha. O tipo de análise que podemos fazer aqui consiste em ver se as proporções de pessoas em cada categoria são as mesmas para as quatro faixas etárias. Ou seja, estamos comparando quatro distribuições multinomiais e isso leva ao *teste de homogeneidade*.

Suponhamos, agora, que uma amostra de n alunos seja retirada e cada aluno seja classificado de acordo com sua faixa etária ***e*** sua opinião sobre o uso medicinal da maconha. Temos, agora, dados emparelhados, ou seja, as amostras são dependentes e o objetivo aqui é ver se há dependência entre as duas variáveis. Isso nos leva ao *teste de independência*.

Em ambos os casos, podemos resumir as informações através de uma tabela de contingência, ou tabela de dupla entrada com as J categorias de uma variável sendo exibidas nas colunas e as I categorias da outra variável nas linhas. Na tabela a seguir ilustra-se a notação a ser utilizada nos testes. O ponto no subscrito indica soma dos valores ao longo da respectiva dimensão.

Tabela 1.2 – Tabela de dados bivariados

		Variável coluna				
		1	2	$\cdots$	J	Total de linha
Variável linha	1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
Total de coluna		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	$n = n_{\cdot\cdot}$

n_{ij} = frequência observada na cela (i, j)

$$n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{iJ} = \sum_{j=1}^{J} n_{ij} \quad \text{(total da linha } i\text{)}$$

$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{Ij} = \sum_{i=1}^{I} n_{ij} \quad \text{(total da coluna } j\text{)}$$

$$n_{\cdot\cdot} = n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \quad \text{(total de observações na amostra)}$$

1.3.1 Amostras independentes: Teste de homogeneidade

Como visto, um dos contextos que origina dados bivariados é quando amostras independentes são retiradas de várias populações e cada indivíduo ou objeto é classificado de acordo com uma variável qualitativa. Neste caso, uma das variáveis é a que identifica a população e a outra é a que classifica os sujeitos. Em tal contexto, nosso interesse é testar se as proporções em cada categoria são as mesmas em todas as populações. No Exemplo 1.2, as populações são formadas pelos homens e pelas mulheres e a variável de classificação é o meio de transporte. O interesse é testar se as proporções de homens e mulheres que usam cada tipo de transporte são iguais. Na primeira situação do exemplo no início da seção, idade é a variável que identifica a população e a opinião sobre o uso medicinal da maconha é a

variável de interesse, que classifica os sujeitos. O objetivo é testar se a proporção de pessoas com cada opinião é a mesma nas quatro faixas etárias.

Na Tabela 1.2, suponhamos que haja J populações com a variável de classificação tendo I categorias. Nosso interesse é testar

$$H_0 : p_{i1} = p_{i2} = \cdots = p_{iJ} \quad \forall i \tag{1.7}$$

Sob H_0, temos, então, que

$$\frac{n_{i1}}{n_{\cdot 1}} = \frac{n_{i2}}{n_{\cdot 2}} = \cdots = \frac{n_{iJ}}{n_{\cdot J}} = \frac{n_{i1} + n_{i2} + \cdots + n_{iJ}}{n_{\cdot 1} + n_{\cdot 2} + \cdots + n_{\cdot J}} = \frac{n_{i\cdot}}{n} = p_{i\cdot}$$

em que $p_{i\cdot}$ é a proporção de elementos na amostra que pertencem à categoria i.

Assim, sob H_0, a proporção na categoria i em cada população deve ser igual à proporção da categoria i na amostra toda. Como há $n_{\cdot j}$ elementos na população j, esperamos, então, que uma proporção $p_{i\cdot}$ desses elementos esteja na categoria i, ou seja,

$$\frac{e_{ij}}{n_{\cdot j}} = p_{i\cdot} = \frac{n_{i\cdot}}{n}$$

Isso nos dá que

$$e_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \tag{1.8}$$

Como antes, a estatística de teste é uma medida de distância relativa entre frequências observadas e frequências esperadas:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{1.9}$$

e se $e_{ij} \geq 5 \quad \forall\, i, j$

$$X^2 \approx \chi^2_{(I-1)(J-1)} \tag{1.10}$$

Valores grandes da estatística levam à rejeição da hipótese nula de igualdade das proporções em cada categoria ao longo das populações.

As mesmas propriedades sobre os valores observados e esperados continuam valendo.

- $\sum_{i=1}^{I} \sum_{j=1}^{J} e_{ij} = n$

  De fato:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} e_{ij} = \frac{1}{n} \sum_{i=1}^{I} n_{i\cdot} \sum_{j=1}^{J} n_{\cdot j} = \frac{1}{n} n \cdot n = n$$

- $\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-e_{ij})=0$

  De fato:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-e_{ij})=\sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij}-\sum_{i=1}^{I}\sum_{j=1}^{J}e_{ij}=n-n=0$$

EXEMPLO 1.6 Meio de transporte e gênero – continuação

Com relação ao Exemplo 1.2, suponhamos que amostras independentes de 300 homens e de 250 mulheres tenham revelado a seguinte distribuição:

Frequências observadas			
Meio de transporte	Homens	Mulheres	Total
Ônibus	77	116	193
Barca	48	19	67
Ônibus e barca	47	54	101
Carro	96	28	124
Caminhada/bicicleta	32	33	65
Total	300	250	550

Vamos realizar o teste de homogeneidade, ou seja, vamos testar, ao nível de significância $\alpha = 0,05$, se as proporções em cada categoria de meio de transporte são iguais entre homens e mulheres. Para isso, precisamos calcular as frequências esperadas em cada cela. Para a cela (1,1) (Homens que vão de ônibus) temos

$$e_{11}=\frac{n_{1\cdot}n_{\cdot 1}}{n}=\frac{193\cdot 300}{550}=105,27273$$

De forma análoga obtemos as frequências esperadas para as outras celas, que estão na tabela a seguir:

Frequências esperadas			
Meio de transporte	Homens	Mulheres	Total
Ônibus	105,27273	87,72727	193
Barca	36,54545	30,45455	67
Ônibus e barca	55,09091	45,90909	101
Carro	67,63636	56,36364	124
Caminhada/bicicleta	35,45455	29,54545	65
Total	300	250	550

Como todas as frequências esperadas são maiores que 5, podemos usar a aproximação qui-quadrado, ou seja, $X^2 \approx \chi^2_4$. Note que o número de graus de liberdade é $(5-1)(2-1)=4$ e $F_{4;0,05}=9,4877$. A parcela da cela (1,1) na estatística qui-quadrado é

$$x_{11}=\frac{(77-105,27273)^2}{105,27273}=7,59311$$

Na tabela a seguir, apresentamos as contribuições de cada cela:

Contribuição para o X^2		
Meio de transporte	Homens	Mulheres
Ônibus	7,59311	9,11173
Barca	3,59023	4,30828
Ônibus e barca	1,18827	1,42592
Carro	11,89443	14,27331
Caminhada/bicicleta	0,3366	0,40392

Somando todas essas celas, obtemos $X^2 = 54,1258 > 9,4877$. Dessa forma, rejeitamos a hipótese nula, ou seja, as proporções em cada categoria de meio de transporte não são iguais entre homens e mulheres. Analisando a tabela das contribuições, podemos ver que as maiores diferenças estão nas categorias de ônibus e carro. Há mais mulheres utilizando ônibus e mais homens utilizando carros do que se esperaria.

Na Figura 1.1 temos a saída do programa Minitab para essa análise.

Estatísticas Tabuladas: MTransp; Gênero

Linhas: MTransp Colunas: Gênero

	Homens	Mulheres	Todos
Ônibus	77	116	193
	105,27	87,73	
	7,593	9,112	
Barca	48	19	67
	36,55	30,45	
	3,590	4,308	
Ônibus&Barca	47	54	101
	55,09	45,91	
	1,188	1,426	
Carro	96	28	124
	67,64	56,36	
	11,894	14,273	
Caminhada/bicicleta	32	33	65
	35,45	29,55	
	0,337	0,404	
Todos	300	250	550

Conteúdo da Célula: Contagem
Contagem esperada
Contribuição para Qui-Quadrado

Qui-Quadrado de Pearson = 54,126; GL = 4; Valor-P = 0,000

Figura 1.1 – saída do Minitab para o Exemplo 1.6

♦♦

1.3.2 Amostras dependentes: Teste de independência

Consideremos, agora, o caso de dados emparelhados, ou seja, cada indivíduo da amostra é classificado segundo duas variáveis qualitativas X e Y. Embora a representação tabular dos dados amostrais seja a mesma apresentada na Tabela 1.2, a interpretação e uso dos dados é completamente diferente. Em cada cela, a contagem n_{ij} nos dá o número de sujeitos para os quais $X = i$ e $Y = j$; nosso interesse é determinar se as variáveis X e Y são independentes. Sabemos que, se X e Y são independentes, então

$$P(X = i, Y = j) = P(X = i)\,P(Y = j)$$

Então, sob a veracidade de H_0, temos que ter

$$\frac{e_{ij}}{n} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}$$

ou equivalentemente

$$e_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \tag{1.11}$$

Se todas as frequências esperadas forem de pelo menos 5, pode-se mostrar que

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \approx \chi^2_{(I-1)(J-1)} \tag{1.12}$$

EXEMPLO 1.7 Meio de transporte e uso do bandejão – continuação

A título de ilustração da diferença de interpretação da estatística qui-quadrado no contexto de dados emparelhados, vamos considerar os mesmos dados do exemplo anterior, mas agora representando uma amostra de 550 alunos, dos quais 300 usam o bandejão e 250 não usam, de acordo com a seguinte distribuição:

Meio de transporte	Usa bandejão	Não usa bandejão	Total
Ônibus	77	116	193
Barca	48	19	67
Ônibus e barca	47	54	101
Carro	96	28	124
Caminhada/bicicleta	32	33	65
Total	300	250	550

O cálculo das frequências esperadas e da estatística de teste é o mesmo e concluímos, ao nível de significância $\alpha = 0,05$, que as variáveis "meio de transporte" e "uso do bandejão" não são independentes.

1.4 Exercícios propostos

1. Em recente estudo, obteve-se uma amostra aleatória de proprietários de pequenas empresas, e pediu-se a cada um que apontasse o maior problema enfrentado por sua empresa. Os resultados são apresentados na seguinte tabela de frequências de uma entrada.

Problema	**Frequência**
Custo do seguro saúde	430
Seguro de responsabilidade	145
Indenização ao trabalhador	135
Custo de combustíveis	90

   Em um relatório econômico, as verdadeiras proporções para cada categoria foram dadas como 0,50, 0,20, 0,20 e 0,10. Esses dados fornecem alguma evidência que contradiga o relatório econômico? Ache limites para o valor P associado a esse teste.

2. Os escritórios de admissão dos campi da Universidade da Califórnia mantêm registros históricos cuidadosos dos candidatos. Em 2008, as proporções de estudantes que se candidataram para as faculdades do sistema universitário por localização na Califórnia foram as seguintes: Los Angeles (LA), 29,2%; San Francisco (SF), 26,1%; Orange County (OC), 9,9%; Riverside/San Bernardino (RS), 7,7%; todas as outras (O), 27,1%.1 Suponha que se tenha obtido uma amostra aleatória de candidatos em 2009 para a qual foram obtidas as seguintes frequências para as localizações.

Localização	LA	SF	OC	RS	O
Frequência	125	96	45	44	128

Há alguma evidência de mudança na proporção de candidatos por localização na Califórnia? Use $\alpha = 0,005$.

3. Embora a harmonização de alimento e vinho seja subjetiva e uma ciência não exata, tradicionalmente diz-se que vinho tinto combina com carne vermelha, e vinho branco combina com peixe e aves. Obteve-se uma amostra aleatória de jantares em restaurantes quatro estrelas, e cada jantar foi classificado de acordo com a comida e o vinho pedido. Eis a tabela de frequências de dupla entrada resultante.

		Vinho	
		Tinto	Branco
Comida	Carne vermelha	86	46
	Peixe ou ave	50	64

Há alguma evidência de que a comida e o vinho sejam dependentes? Teste a hipótese relevante com $\alpha = 0,005$. Esses dados sugerem que os jantares ainda estejam seguindo as harmonizações tradicionais de comida e vinho?

4. Obtiveram-se amostras aleatórias de apostadores em quatro cassinos de Las Vegas, e perguntou-se a cada apostador qual jogo jogava mais. Os resultados são apresentados na tabela de frequências de dupla entrada que segue.

		Jogo			
		Blackjack	Pôquer	Roleta	Caça-níquel
Cassino	Bellagio	22	20	38	66
	Caesar's	30	38	22	68
	Golden Nugget	28	25	21	81
	Harrah's	38	25	29	84

Realize um teste para homogeneidade de populações. Há alguma evidência que sugira que a verdadeira proporção de apostadores em cada jogo não seja a mesma para todos os cassinos? Use $\alpha = 0,05$.

5. Foram obtidas amostras aleatórias de clientes de duas lojas diferentes de material para escritório, e perguntou-se a cada cliente qual tipo de recurso para escrita preferiam. Os resultados estão resumidos na seguinte tabela de frequências de dupla entrada.

		Recurso para escrita		
		Lápis	Lapiseira	Caneta
Loja	Casa Cruz	183	164	480
	Kalunga	130	202	420

Realize um teste para homogeneidade de populações. Use $\alpha = 0,01$. Estabeleça sua conclusão, justifique sua resposta e ache limites para o valor P associado a esse teste.

1.5 Solução dos exercícios propostos

1. Teste de aderência

   $p_{10} = 0,5 \qquad p_{20} = 0,2 \qquad p_{30} = 0,2 \qquad p_{40} = 0,1$

   $$H_0 : p_i = p_{i0} \qquad \forall i$$
   $$H_1 : \text{pelo menos um } p_i \neq p_{i0}$$

   Sob H_0, as frequências esperadas são

   $$E_1 = 0,5 \times 800 = 400 \qquad E_2 = 0,2 \times 800 = 160$$
   $$E_3 = 0,2 \times 800 = 160 \qquad E_4 = 0,1 \times 800 = 80$$

   $E_i \geq 5 \quad \forall i$ – aproximação qui-quadrado OK! $\qquad k = 4 \rightarrow \chi^2_{3;0,05} = 7,815$

   Valor observado da estatística de teste:

   $$\chi^2_0 = \frac{(430-400)^2}{400} + \frac{(145-160)^2}{160} + \frac{(135-160)^2}{160} + \frac{(90-80)^2}{80} = 8,8125$$

   Rejeita-se H_0; há evidências, ao nível de 5%, de que pelo menos uma das proporções p_i é diferente do valor dado no relatório. Veja saída do Minitab na Figura 1.2.

Teste Qui-Quadrado de Qualidade de Ajuste para Contagens Observadas na Variável: Problema

Categoria	Observado	Teste de Proporção	Esperado	Contribuição para Qui-Quadrado
SeguroSaúde	430	0,5	400	2,25000
SeguroResp	145	0,2	160	1,40625
Indenização	135	0,2	160	3,90625
Combustíveis	90	0,1	80	1,25000

N	GL	Qui-Quadrado	Valor-p
800	3	8,8125	0,032

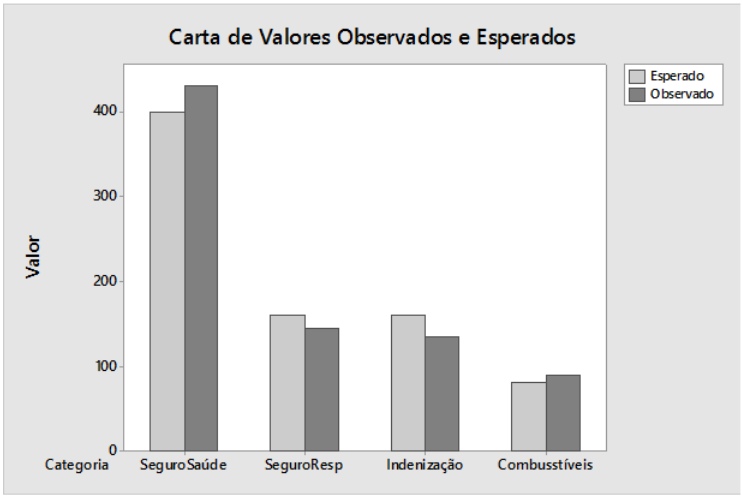

Figura 1.2 – Saída do Minitab para o Exercício 1

2. Teste de aderência

   $p_{LA,0} = 0,292 \qquad p_{SF,0} = 0,261 \qquad p_{OC,0} = 0,099 \qquad p_{RS,0} = 0,077 \qquad p_{O,0} = 0,271$

   $$H_0 : p_i = p_{i0} \qquad \forall i$$
   $$H_1 : \text{pelo menos um } p_i \neq p_{i0}$$

Sob H_0, as frequências esperadas são

$$E_{LA} = 0,292 \times 438 = 127,896 \qquad E_{SF} = 0,261 \times 438 = 114,318$$
$$E_{OC} = 0,099 \times 438 = 43,362 \qquad E_{RS} = 0,077 \times 438 = 33,726 \qquad E_O = 0,271 \times 438 = 118,698$$

$E_i \geq 5 \quad \forall i$ – aproximação qui-quadrado OK! $\qquad k = 5 \rightarrow \chi^2_{4;0,005} = 14,860$

Valor observado da estatística de teste:

$$\chi^2_0 = \frac{(125-127,896)^2}{127,896} + \frac{(96-114,318)^2}{114,318} + \frac{(45-43,362)^2}{43,362} + \frac{(44-33,726)^2}{33,726} + \frac{(128-118,698)^2}{118,698} = 6,92143$$

Não se rejeita H_0; não há evidências, ao nível de 0,5%, de que as proporções p_i em 2009 sejam diferentes das proporções em 2008. Veja saída do Minitab na Figura 1.3.

Teste Qui-Quadrado de Qualidade de Ajuste para Contagens Observadas na Variável: Local

Categoria	Observado	Teste de Proporção	Esperado	Contribuição para Qui-Quadrado
LA	125	0,292	127,896	0,06558
SF	96	0,261	114,318	2,93523
OC	45	0,099	43,362	0,06188
RS	44	0,077	33,726	3,12978
O	128	0,271	118,698	0,72897

N	GL	Qui-Quadrado	Valor-p
438	4	6,92143	0,140

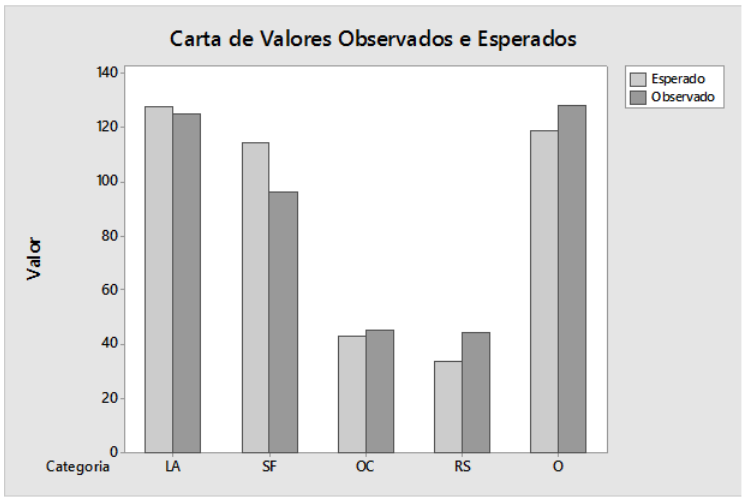

Figura 1.3 – Saída do Minitab para o Exercício 2

3. Dados emparelhados - teste de independência

H_0 : vinho e carne são independentes
H_1 : vinho e carne não são independentes

Sob H_0, as frequências esperadas são

$$e_{11} = \frac{132 \cdot 136}{246} = 72,9756 \qquad e_{12} = \frac{132 \cdot 110}{246} = 59,0244$$
$$e_{21} = \frac{114 \cdot 136}{246} = 63,0244 \qquad e_{12} = \frac{114 \cdot 110}{246} = 50,9756$$

$E_i \geq 5 \quad \forall i$ – aproximação qui-quadrado OK! $gl = (2-1)(2-1) = 1 \rightarrow \chi^2_{1;0,005} = 7,879$ Valor observado da estatística de teste:

$$\chi^2_0 = \frac{(86-72,9756)^2}{72,9756} + \frac{(46-59,0244)^2}{59,0244} + \frac{(50-63,0244)^2}{63,0244} + \frac{(64-50,9756)^2}{50,9756} = 11,2178$$

Rejeita-se H_0; há evidências, ao nível de 0,5%, de que as variáveis "escolha do vinho" e "escolha da carne" não são independentes. Veja saída do Minitab na Figura 1.4.

Teste Qui-Quadrado para Independência: C13; Vinho

```
Linhas: Carne  Colunas: Vinho

               Tinto   Branco  Todos

Vermelha          86       46    132
                72,98    59,02
                2,325    2,874

Branca            50       64    114
                63,02    50,98
                2,692    3,328

Todos            136      110    246

Conteúdo da Célula:      Contagem
                         Contagem esperada
                         Contribuição para Qui-Quadrado


Qui-Quadrado de Pearson = 11,218; GL = 1; Valor-P = 0,001
Qui-Quadrado da Razão de Verossimilhanças = 11,284; GL = 1; Valor-P = 0,001
```

Figura 1.4 – Saída do Minitab para o Exercício 3

4. Teste de homogeneidade

O objetivo é testar se as proporções de usuários de cda tipo de recurso de escrita são as mesmas para as 2 lojas. Na Figura 1.6 temos a saída o Minitab. Aí podemos ver que todas as frequências esperadas são maiores que 5, o que permite o uso da aproximação qui-quadrado. O número de graus de liberdade é $gl = (2-1)(3-1) = 2$ e o valor crítico para $\alpha = 0,01$ é 9,210. O valor da estatística de teste é 13,388, o que nos leva à rejeição da hipótese nula, ou seja, as proporções de usuários dos diferentes tipos de recursos não são as mesmas nas 2 lojas. Pela tabela, temos que o valor P está entre 0,001 e 0,005. O Minitab dá $P = 0,001$, mas o valor mais preciso é 0,001238.

5. Teste de homogeneidade

O objetivo é testar se as proporções de jogadores em cada um dos jogos são as mesmas nos 4 cassinos. Na Figura 1.5 temos a saída o Minitab. Aí podemos ver que todas as frequências esperadas são maiores que 5, o que permite o uso da aproximação qui-quadrado. O número de graus de liberdade é $gl = (4-1)(4-1) = 9$ e o valor crítico para $\alpha = 0,05$ é 16,919. O valor da estatística de teste é 18,801, o que nos leva à rejeição da hipótese nula, ou seja, as proporções de jogadores nos diferentes jogos não são as mesmas em todos os cassinos. Pela tabela, temos que o valor P está entre 0,025 e 0,05. O Minitab dá $P = 0,027$.

Teste Qui-Quadrado para Homogeneidade: Jogo; Cassino

Linhas: Jogo Colunas: Cassino

	Bellagio	Caesar	GoldenNugget	Harrahs	Todos
Blackjack	22	30	28	38	118
	27,00	29,22	28,67	33,11	
	0,9270	0,0207	0,0156	0,7233	
Poquer	20	38	25	25	108
	24,71	26,75	26,24	30,30	
	0,8994	**4,7353**	0,0584	0,9274	
Roleta	38	22	21	29	110
	25,17	27,24	26,72	30,86	
	6,5368	1,0085	1,2261	0,1123	
CacaNiquel	66	68	81	87	302
	69,11	74,79	73,37	84,73	
	0,1399	0,6164	0,7935	0,0608	
Todos	146	158	155	179	638

Conteúdo da Célula: Contagem
Contagem esperada
Contribuição para Qui-Quadrado

Qui-Quadrado de Pearson = 18,801; GL = 9; Valor-P = 0,027
Qui-Quadrado da Razão de Verossimilhanças = 17,678; GL = 9; Valor-P = 0,039

Figura 1.5 – Saída do Minitab para o Exercício 4

Teste Qui-Quadrado para Homogeneidade: Recurso; Loja

Linhas: Recurso Colunas: Loja

	CCruz	Kalunga	Todos
Lápis	183	130	313
	163,9	149,1	
	2,218	2,439	
Lapiseira	164	202	366
	191,7	174,3	
	4,000	4,399	
Caneta	480	420	900
	471,4	428,6	
	0,158	0,174	
Todos	827	752	1579

Conteúdo da Célula: Contagem
Contagem esperada
Contribuição para Qui-Quadrado

Qui-Quadrado de Pearson = 13,388; GL = 2; Valor-P = 0,001
Qui-Quadrado da Razão de Verossimilhanças = 13,410; GL = 2; Valor-P = 0,001

Figura 1.6 – Saída do Minitab para o Exercício 5