Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-04-06

Sumário

Prefácio

Capítulo 1

Introdução

Capítulo 2

Referencial para Inferência

Capítulo 3

Estimação Baseada no Plano Amostral

Capítulo 4

Efeitos do Plano Amostral

Capítulo 5

Ajuste de Modelos Paramétricos

Capítulo 6

Modelos de Regressão

Capítulo 7

Testes de Qualidade de Ajuste

Capítulo 8

Testes em Tabelas de Duas Entradas

8.1 Introdução

Os principais testes em tabelas de duas entradas são os de homogeneidade e de independência. O `teste de homogeneidade` é apropriado para estudar a igualdade das distribuições condicionais de uma variável resposta categórica correspondentes a diferentes níveis de uma variável preditora também categórica. O `teste de independência` é adequado para estudar a associação entre duas variáveis categóricas. Enquanto o primeiro teste se refere às distribuições condicionais da variável resposta para níveis fixados da variável preditora, o segundo se refere à distribuição conjunta das duas variáveis categóricas que definem as celas da tabela. Apesar de conceitualmente distintas, as duas hipóteses podem ser testadas, no caso de amostragem aleatória simples, utilizando a mesma estatística de teste multinomial de Pearson.

Nos testes de homogeneidade e de independência para tabelas de frequências $L \times C$ obtidas por amostragem aleatória simples, a estatística de teste de Pearson tem distribuição assintótica qui-quadrado com $(L-1)(C-1)$ graus de liberdade, isto é $\chi^2\left((L-1)(C-1)\right)$. Para pesquisas com planos amostrais complexos, esta propriedade assintótica padrão não é válida. Por exemplo, testes definidos em tabelas de frequências obtidas mediante amostragem por conglomerados são mais liberais (rejeitam mais) relativamente aos níveis nominais de significância, devido à correlação intraclasse positiva das variáveis usadas para definir a tabela. Além disso, para planos amostrais complexos, as estatísticas de teste das duas hipóteses devem ser corrigidas de formas diferentes.

Neste capítulo, apresentamos versões modificadas de procedimentos clássicos de testes para dados categóricos, de maneira a incorporar os efeitos de plano amostral na análise. Procedimentos mais recentes, baseados em ajustes de modelos regressivos, estão disponíveis em pacotes especializados como o SUDAAN (procedimento CATAN, para dados tabelados, e procedimento LOGISTIC, para regressão com respostas individuais binárias, por exemplo), porém não serão aqui considerados.

8.2 Tabelas 2x2

Para fixar idéias, vamos considerar inicialmente uma tabela de contingência 2×2, isto é, com $L = 2$ e $C = 2$, representada pela Tabela 8.1. A entrada p_{lc} na Tabela 8.1 representa a proporção populacional de unidades no nível l da variável 1 e c da variável 2, ou seja $p_{lc} = \frac{N_{lc}}{N}$, onde N_{lc} é o número de observações na cela (l, c) na população, N é o tamanho da população e $\sum_l \sum_c p_{lc} = 1$. Vamos denotar, ainda, as proporções marginais na tabela por $p_{l+} = \sum_c p_{lc}$ e $p_{+c} = \sum_l p_{lc}$.

Tabela 8.1: Tabela 2x2 de proporções

	Var 2		
Var 1	1	2	Total
1	p_{11}	p_{12}	p_{1+}
2	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

8.2.1 Teste de Independência

A hipótese de independência corresponde a

$$H_0 : p_{lc} = p_{l+}p_{+c} \ \ \forall l, c = 1, 2 \ .$$

A estatística de teste de Pearson para testar esta hipótese, no caso de amostragem aleatória simples, é dada por

$$X_P^2 (I) = n \sum_{l=1}^{2} \sum_{c=1}^{2} \frac{\left(\hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c}\right)^2}{\hat{p}_{l+}\hat{p}_{+c}}$$

onde $\hat{p}_{lc} = n_{lc}/n$, n_{lc} é o número de observações da amostra na cela (l, c) da tabela, n é o tamanho total da amostra, $\hat{p}_{l+} = \sum_c \widehat{p}_{lc}$ e $\hat{p}_{+c} = \sum_l \hat{p}_{lc}$.

Sob a hipótese nula, a estatística $X_P^2 (I)$ tem distribuição de referência qui-quadrado com um grau de liberdade. Observe que esta estatística mede uma distância (em certa escala) entre os valores observados na amostra e os valores esperados (estimados) sob a hipótese nula de independência.

8.2.2 Teste de Homogeneidade

No caso do teste de independência, as duas variáveis envolvidas são consideradas como respostas. No teste de homogeneidade, uma das variáveis, a variável 2, por exemplo, é considerada a resposta enquanto a variável 1 é considerada explicativa. Vamos agora analisar a distribuição da variável 2 (coluna) para cada nível da variável 1 (linha). Considerando ainda uma tabela 2×2, queremos testar a hipótese

$$H_0 : p_{1c} = p_{2c} \quad c = 1, 2 \ .$$

onde agora p_{lc} representa a proporção na linha l de unidades na coluna c. Com as restrições usuais de que as proporções nas linhas somam 1, isto é, $p_{11} + p_{12} = p_{21} + p_{22} = 1$, a hipótese nula considerada se reduz a $p_{11} = p_{21}$ e novamente temos apenas um grau de liberdade.

Para o teste de homogeneidade, usamos a seguinte estatística de teste de Pearson:

$$X_P^2 (H) = \sum_{l=1}^{2} \sum_{c=1}^{2} \frac{n_{l+} \left(\widehat{p}_{lc} - \hat{p}_{+c}\right)^2}{\hat{p}_{+c}},$$

onde $n_{l+} = \sum_c n_{lc}$ para $l = 1, 2$ e $\widehat{p}_{lc} = n_{lc}/n_{l+}$ para $l = 1, 2$ e $c = 1, 2$.

Esta estatística mede a distância entre valores observados e esperados sob a hipótese nula de homogeneidade e tem, também, distribuição de referência qui-quadrado com um grau de liberdade. Embora as expressões de $X_P^2 (I)$ e $X_P^2 (H)$ sejam distintas, seus valores numéricos são iguais.

8.2.3 Efeitos de Plano Amostral nas Celas

Para relacionar os testes tratados neste capítulo com o teste de qualidade de ajuste apresentado no capítulo anterior, observe que os testes de independência e de homogeneidade são definidos sobre o vetor de proporções de distribuições multinomiais. No caso de independência, temos uma distribuição multinomial com vetor de probabilidades $(p_{11}, p_{12}, p_{21}, p_{22})$, e no caso do teste de homogeneidade, temos duas multinomiais (no caso binomiais) com vetores de probabilidades (p_{11}, p_{12}) e (p_{21}, p_{22}). O processo de contagem que gera estas multinomiais pressupõe que as observações individuais (indicadores de classe) são independentes e com mesma distribuição. Estas hipóteses só são válidas no caso de amostragem aleatória simples com reposição.

Quando os dados são gerados através de um plano amostral complexo, surgem efeitos de conglomeração e estratificação que devem ser considerados no cálculo das estatísticas de teste. Neste caso, as frequências nas celas da tabela são estimadas, levando em conta os pesos dos elementos da amostra bem como o plano amostral efetivamente utilizado.

Denotemos por $\hat{N}_{lc}$ o estimador do número de observações na cela (l, c) na população, e designemos por $\hat{n}_{lc} = \left(\hat{N}_{lc}/\hat{N}\right) \times n$ o valor padronizado de $\hat{N}_{lc}$, de modo que $\sum_{l=1}^{L}\sum_{c=1}^{C} \hat{n}_{lc} = n$. Sejam, agora, os estimadores das proporções nas celas dados por $\hat{p}_{lc} = \hat{n}_{lc}/n$ no caso do teste de independência e por $\hat{p}_{lc} = \hat{n}_{lc}/n_{l+}$ no caso do teste de homogeneidade. As estatísticas $X^2_P(I)$ e $X^2_P(H)$ calculadas com as estimativas $\hat{n}_{lc}$ no lugar dos valores n_{lc} não têm, como antes, distribuição assintótica qui-quadrado com um grau de liberdade.

Por outro lado, é importante observar que as agências produtoras de dados estatísticos geralmente apresentam os resultados de suas pesquisas em tabelas contendo as estimativas $\hat{N}_{lc}$, como ilustrado no Exemplo **??** do Capítulo 5. Se calcularmos as estatísticas $X^2_P(I)$ e $X^2_P(H)$ a partir dos valores dos $\hat{N}_{lc}$ fornecidos, com a estimativa do tamanho da população $\hat{N}$ no lugar de n, os resultados assintóticos obtidos para amostragem aleatória simples com reposição (IID) deixarão de ser válidos. Devemos calcular as estatísticas de teste $X^2_P(I)$ e $X^2_P(H)$ a partir dos $\hat{n}_{lc}$ anteriormente definidos, que correspondem aos $\hat{N}_{lc}$ padronizados para totalizar n.

As estatísticas baseadas nos valores estimados $\hat{n}_{lc}$ podem ser corrigidas para ter distribuição de referência qui-quadrado com um grau de liberdade, no caso de tabela 2×2. Mas, é importante observar que os efeitos de plano amostral e as correções a serem considerados são distintos para as duas estatísticas $X^2_P(I)$ e $X^2_P(H)$.

Para ilustrar esse ponto vamos considerar o ajuste de EPA médio, que será apresentado na próxima seção para o caso de tabelas $L \times C$. Este ajuste, no caso da estatística $X^2_P(I)$, se baseia no EPA médio das estimativas das proporções nas celas $\hat{p}_{lc} = \hat{n}_{lc}/n$, enquanto que para a estatística $X^2_P(H)$ ele se baseia no EPA médio das estimativas das proporções nas linhas $\hat{p}_{lc} = \hat{n}_{lc}/n_{l+}$.

Os valores das estatísticas $X^2_P(I)$ e $X^2_P(H)$ são iguais no caso IID, mas para planos amostrais complexos, as estatísticas corrigidas pelo EPA médio são distintas, apesar de terem, para tabelas 2×2, a mesma distribuição de referência qui- quadrado com um grau de liberdade. Adiante apresentaremos um exemplo numérico para ilustrar este ponto.

8.3 Tabelas de Duas Entradas (Caso Geral)

8.3.1 Teste de Homogeneidade

O teste de homogeneidade pode ser usado para comparar distribuições de uma variável categórica (C categorias) para um conjunto de L regiões não superpostas, a partir de amostras independentes obtidas através de um plano amostral com vários estágios. Vamos considerar uma tabela $L \times C$ e supor que as colunas da tabela correspondem às classes da variável resposta e as linhas correspondem às regiões, de modo que as somas da proporções nas linhas na tabela de proporções são iguais a 1. A tabela para a população é da forma da Tabela 8.2.

Tabela 8.2: Proporções de linhas em tabela $L \times C$

Região	1	2	...	c	...	C	Total
1	p_{11}	p_{12}	...	p_{1c}	...	p_{1C}	1
2	p_{21}	p_{22}	...	p_{2c}	...	p_{2C}	1
⋮	⋮	⋮	.	⋮	.	⋮	⋮
l	p_{l1}	p_{l1}	...	p_{lc}	...	p_{lC}	1
⋮	⋮	⋮	.	⋮	.	⋮	⋮
L	p_{L1}	p_{L2}	...	p_{Lc}	...	p_{LC}	1

Note que aqui as proporções que aparecem nas linhas da tabela são proporções calculadas em relação à freqüência total da linha, e não proporções calculadas em relação ao total da tabela como na seção anterior. Portanto, $p_{lc} = N_{lc}/N_{l+}$ para todo $l = 1, \ldots, L$ e $c = 1, \ldots, C$.

Vamos considerar o caso em que $L = 2$ regiões devem ser comparadas. Seja $\mathbf{p}_l = (p_{l1}, \ldots, p_{l\,C-1})'$ o vetor de proporções da l-ésima região, sem incluir a proporção referente à última categoria (p_{lC}), $l = 1, 2$. A hipótese de igualdade das distribuições da resposta nas duas regiões pode ser expressa como $H_0 : \mathbf{p}_1 = \mathbf{p}_2$, com $C-1$ componentes em cada vetor, pois em cada região a soma das proporções é 1.

Seja $\mathbf{p}_0 = (p_{+1}, \ldots p_{+\,C-1})'$ o vetor comum de proporções sob H_0, desconhecido. Denotemos por $\hat{\mathbf{p}}_l = (\hat{p}_{l1}, \ldots, \hat{p}_{l\,C-1})'$ os vetores de proporções estimadas ($l = 1, 2$), baseados em amostras independentes para as diferentes regiões, onde $\hat{p}_{lc} = \widehat{N}_{lc}/\widehat{N}_{l+}$ é um estimador consistente da proporção p_{lc} na população correspondente, e $\widehat{N}_{lc}$ e $\widehat{N}_{l+}$ são estimadores ponderados das frequências nas celas e nas marginais de linha da tabela, respectivamente, de modo que $\sum_{c=1}^{C} \widehat{N}_{lc} = \widehat{N}_{l+}$. Estes estimadores levam em consideração as probabilidades desiguais de inclusão na amostra e os ajustes por não-resposta. Observe que, se os tamanhos das amostras dos subgrupos regionais não forem fixados, os $\hat{p}_{lc}$ são estimadores de razão.

Sejam $\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_1)$ e $\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_2)$ estimadores consistentes das matrizes de variância de aleatorização dos vetores $\widehat{\mathbf{p}}_1$ e $\widehat{\mathbf{p}}_2$, respectivamente. A estatística de Wald baseada no plano amostral $X^2_W(H)$ para efetuar o teste de homogeneidade no caso de duas regiões ($L = 2$) é dada por

$$X^2_W(H) = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)' \left[\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_1) + \hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_2)\right]^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2), \tag{8.1}$$

pois as amostras são disjuntas e supostas independentes.

No caso, a estatística de Wald $X^2_W(H)$ tem distribuição assintótica qui-quadrado com $(2-1)\times(C-1)$ graus de liberdade. Quando o número de unidades primárias de amostragem na amostra de cada região é grande, a estatística de Wald funciona adequadamente. Caso contrário, ocorre problema de instabilidade e usamos, alternativamente, uma estatística F-corrigida de Wald. Freitas et al.(1997) descrevem uma aplicação da estatística $X^2_W(H)$ para testar a hipótese de igualdade das pirâmides etárias estimadas pela Pesquisa sobre Padrões de Vida 96/97 (PPV) e da Pesquisa Nacional por Amostra de Domicílios 95 para as regiões Sudeste e Nordeste. Tal comparação fez parte do processo de avaliação da qualidade dos resultados da PPV.

Designemos por $f = m-H$ o número total de graus de liberdade disponível para estimar $\left[\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_1) + \hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_2)\right]$, onde m e H são os números totais de conglomerados e de estratos nas amostras das duas regiões, respectivamente. As correções F da estatística $X^2_W(H)$ são dadas por

$$F_{1.p} = \frac{f-(C-1)+1}{f(C-1)} X^2_W(H), \tag{8.2}$$

que tem distribuição de referência F com $(C-1)$ e $(f-(C-1)+1)$ graus de liberdade e, ainda,

$$F_{2.p} = X^2_W(H) / (C-1) \tag{8.3}$$

que tem distribuição de referência F com $(C-1)$ e f graus de liberdade.

As estatísticas $F_{1.p}$ e $F_{2.p}$ podem amenizar o efeito de instabilidade, quando f não é grande relativamente ao número de classes (C) da variável resposta.

No caso de $L = 2$ regiões, a estatística de teste de homogeneidade de Pearson é dada por

$$X^2_P(H) = (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2)' \left(\hat{\mathbf{P}}/\widehat{n}_{1+} + \hat{\mathbf{P}}/\widehat{n}_{2+}\right)^{-1} (\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2) \quad , \tag{8.4}$$

onde $\hat{\mathbf{P}} = \mathbf{diag}(\hat{\mathbf{p}}_0) - \hat{\mathbf{p}}_0\hat{\mathbf{p}}_0'$ e $\hat{\mathbf{p}}_0$ é o estimador do vetor comum de proporções sob a hipótese de homogeneidade.

Neste caso, $\hat{\mathbf{P}}/\widehat{n}_{1+}$ é o estimador da matriz de covariância de $\hat{\mathbf{p}}_0$ na primeira região e $\hat{\mathbf{P}}/\widehat{n}_{2+}$ na segunda. Observe que (8.4) e (8.1) têm a mesma forma, diferindo só no estimador da matriz de covariância usado para definir a métrica de distância. No caso da estatística $X^2_P(H)$, o estimador da matriz de covariância baseia-se nas hipóteses relativas à distribuição multinomial, apropriadas para a amostragem aleatória simples. A distribuição de referência da estatística $X^2_P(H)$ é qui-quadrado com $(C-1)$ graus de liberdade.

Para introduzir em $X^2_P(H)$ o ajuste de EPA médio e o ajuste de Rao-Scott de primeira ordem, é preciso calcular estimativas de efeitos de plano amostral das estimativas das proporções nas linhas em ambas as regiões. O ajuste de segunda ordem de Rao-Scott, por sua vez, depende da matriz de efeito multivariado do plano amostral. As estimativas de efeitos de plano amostral na região l são da forma

$$\hat{d}_{lc} = \widehat{n}_{l+}\hat{V}_{lc} / \left(\hat{p}_{+c}(1-\hat{p}_{+c})\right), \quad l = 1,2 \text{ e } c = 1,\ldots,C, \tag{8.5}$$

onde $\hat{V}_{lc}$ é o c-ésimo elemento da diagonal de $\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_l)$.

A matriz estimada de efeito multivariado de plano amostral é

$$\hat{\mathbf{\Delta}} = \frac{\widehat{n}_{1+} \times \widehat{n}_{2+}}{\widehat{n}_{1+} + \widehat{n}_{2+}} \hat{\mathbf{P}}^{-1} \left(\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_1) + \hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_2)\right) \quad . \tag{8.6}$$

A estatística de Pearson com ajuste de EPA médio é dada por

$$X^2_P\left(H; \hat{d}_\cdot\right) = X^2_P(H) / \hat{d}_\cdot, \tag{8.7}$$

onde $\hat{d}_\cdot = \sum_{l=1}^{2}\sum_{c=1}^{C} \hat{d}_{lc}/2C$ é a média das estimativas dos efeitos univariados de plano amostral.

Usando os autovalores $\hat{\delta}_c$ de $\hat{\mathbf{\Delta}}$, o ajuste de primeira ordem de Rao-Scott é dado por

$$X^2_P\left(H; \hat{\delta}_\cdot\right) = X^2_P(H) / \hat{\delta}_\cdot, \tag{8.8}$$

onde

$$\hat{\delta}_\cdot = \frac{tr\left(\hat{\mathbf{\Delta}}\right)}{(C-1)} = \frac{1}{C-1}\sum_{l=1}^{2}\left(1 - \frac{\widehat{n}_{l+}}{\widehat{n}_{1+} + \widehat{n}_{2+}}\right)\sum_{c=1}^{C}\frac{\hat{p}_{lc}}{\hat{p}_{+c}}(1-\hat{p}_{lc})\,\hat{d}_{lc}$$

é um estimador da média $\bar{\delta}$ dos autovalores δ_c da matriz $\mathbf{\Delta}$, desconhecida, de efeito multivariado do plano amostral. Como a soma dos autovalores de $\hat{\mathbf{\Delta}}$ é igual ao traço de $\hat{\mathbf{\Delta}}$, esta correção pode ser obtida sem ser necessário calcular os autovalores.

Tabela 8.3: Proporções por cela na população

	Variável 2						
Variável 1	1	2	...	c	...	C	Total
1	p_{11}	p_{12}	...	p_{1c}	...	p_{1C}	p_{1+}
2	p_{21}	p_{22}	...	p_{2c}	...	p_{2C}	p_{2+}
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
l	p_{l1}	p_{l1}	...	p_{lc}	...	p_{lC}	p_{l+}
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
L	p_{L1}	p_{L2}	...	p_{Lc}	...	p_{LC}	p_{L+}
Total	p_{+1}	p_{+2}	...	p_{+c}	...	p_{+C}	1

As distribuições de referência, tanto de $X_P^2\left(H;\hat{d}_{\cdot}\right)$ como de $X_P^2\left(H;\hat{\delta}_{\cdot}\right)$, são qui-quadrado com $(C-1)$ graus de liberdade. Estes ajustes corrigem a estatística $X_P^2(H)$ de modo a obter estatísticas com valor esperado igual ao da distribuição qui-quadrado de referência. Tal correção é apropriada quando houver pouca variação das estimativas dos autovalores $\hat{\delta}_c$. Quando isto não ocorrer, pode ser introduzido o ajuste de segunda ordem de Rao-Scott, que para a estatística de Pearson é dado por

$$X_P^2\left(H;\hat{\delta}_{\cdot},\hat{a}^2\right) = X_P^2\left(H;\hat{\delta}_{\cdot}\right)/\left(1+\hat{a}^2\right) \tag{8.9}$$

onde $\hat{a}^2$ é o quadrado do coeficiente de variação dos quadrados das estimativas dos autovalores $\hat{\delta}_c$, dado por

$$\hat{a}^2 = \sum_{c=1}^{C} \hat{\delta}_c^2 / \left((C-1)\,\hat{\delta}_{\cdot}^2\right) - 1,$$

onde a soma dos quadrados dos autovalores pode ser obtida a partir do traço de $\hat{\mathbf{\Delta}}^2$

$$\sum_{c=1}^{C} \hat{\delta}_c^2 = tr\left(\hat{\mathbf{\Delta}}^2\right) \quad .$$

A estatística de Pearson com a correção de segunda ordem de Rao-Scott $X_P^2\left(H;\hat{\delta}_{\cdot},\hat{a}^2\right)$ tem distribuição de referência qui-quadrado com graus de liberdade com ajuste de Satterhwaite $gl_S = (C-1)/\left(1+\hat{a}^2\right)$.

Quando as estimativas $\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_1)$ e $\hat{\mathbf{V}}_p(\widehat{\mathbf{p}}_2)$ das matrizes de covariâncias regionais são baseadas em números relativamente pequenos de unidades primárias de amostragem selecionadas, pode-se usar a estatística F-corrigida de Pearson. Ela é dada, no caso de duas regiões, por

$$FX_P^2\left(H;\hat{\delta}_{\cdot}\right) = X_P^2\left(H;\hat{\delta}_{\cdot}\right)/(C-1),$$

e tem distribuição de referência F com $(C-1)$ e f graus de liberdade.

8.3.2 Teste de Independência

Vamos considerar o teste de independência no caso geral de tabela $L \times C$, onde os dados são extraídos de uma única população, sem fixar marginais. Consideremos a Tabela 8.3 com as proporções nas celas a nível da população, onde agora novamente se tem $p_{lc} = N_{lc}/N$.

Estamos interessados em testar a hipótese de independência

$$H_0 : p_{lc} = p_{l+}p_{+c}, \quad l = 1, \ldots, L-1,\ c = 1, \ldots, C-1,$$

onde $p_{l+} = \sum_{c=1}^{C} p_{lc}$, $p_{+c} = \sum_{l=1}^{L} p_{lc}$ e $\sum_{c=1}^{C}\sum_{l=1}^{L} p_{lc} = 1$.

Vamos escrever a hipótese de independência numa forma alternativa mas equivalente, usando contrastes de proporções:

$$H_0 : f_{lc} = p_{lc} - p_{l+}p_{+c} = 0, \quad l = 1, \ldots, L-1,\ c = 1, \ldots, C-1.$$

Consideremos o vetor $\mathbf{f}$ com $(L-1)(C-1)$ componentes formado pelos contrastes f_{lc} arranjados em ordem de linhas:

$$\mathbf{f} = (f_{11}, \ldots, f_{1\,C-1}, \ldots, f_{L-1\,1}, \ldots, f_{L-1\,C-1})' .$$

Um teste da hipótese de independência pode ser definido em termos da distância entre uma estimativa consistente do vetor de contrastes $\mathbf{f}$ e o vetor nulo com mesmo número de componentes. O vetor de estimativa consistente de $\mathbf{f}$ é denotado por $\hat{\mathbf{f}} = \left(\hat{f}_{11}, \ldots, \hat{f}_{1\,C-1}, \ldots, \hat{f}_{L-1\,1}, \ldots, \hat{f}_{L-1\,C-1}\right)'$, onde $\hat{f}_{lc} = \hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c}$, onde $\hat{p}_{lc} = \hat{n}_{lc}/n$. Os $\hat{n}_{lc}$ são as frequências ponderadas nas celas, considerando as diferentes probabilidades de inclusão e ajustes por não-resposta, onde os pesos amostrais são normalizados de modo que $\sum_{c=1}^{C}\sum_{l=1}^{L}\hat{n}_{lc} = n$. Se n não for fixado de antemão, os $\hat{p}_{lc}$ serão estimadores de razões. Apenas $(L-1)(C-1)$ componentes são incluídos no vetores $\mathbf{f}$ e $\hat{\mathbf{f}}$, pois a soma das proporções nas celas da tabela é igual a 1.

8.3.3 Estatística de Wald Baseada no Plano Amostral

A estatística de Wald baseada no plano amostral $X^2_W(I)$, para o teste de independência, tem a forma da expressão (8.8), com $\hat{\mathbf{f}}$ no lugar de $\hat{\mathbf{p}}$, o vetor $\mathbf{0}_{(L-1)(C-1)}$ no lugar de $\mathbf{p}_0$ e a estimativa baseada no plano amostral $\hat{\mathbf{V}}_{\mathbf{f}}$ da matriz de covariância de $\hat{\mathbf{f}}$ no lugar de $\hat{\mathbf{V}}_p$. Assim, a estatística de teste de independência de Wald é dada por

$$X^2_W(I) = \hat{\mathbf{f}}'\hat{\mathbf{V}}_{\mathbf{f}}^{-1}\hat{\mathbf{f}} \ , \tag{8.10}$$

que é assintoticamente $\chi^2((L-1)(C-1))$.

A estimativa $\hat{\mathbf{V}}_{\mathbf{f}}$ da matriz de covariância de $\hat{\mathbf{f}}$ pode ser obtida pelo método de linearização de Taylor apresentado na Seção **??**, considerando o vetor de contrastes $\mathbf{f}$ como uma função (não-linear) do vetor $\mathbf{p}$, isto é, $\mathbf{f} = \mathbf{g}(\mathbf{p}) = \mathbf{g}(p_{11}, \ldots, p_{1\,C-1}, \ldots, p_{L-1\,1}, \ldots, p_{L-1\,C-1})$. Assim, a matriz de covariância de $\hat{\mathbf{f}}$ pode ser estimada por

$$\hat{\mathbf{V}}_{\mathbf{f}} = \boldsymbol{\Delta}\mathbf{g}(\hat{\mathbf{p}})\,\hat{\mathbf{V}}_p^{-1}\boldsymbol{\Delta}\mathbf{g}(\hat{\mathbf{p}})' , \tag{8.11}$$

onde $\boldsymbol{\Delta}\mathbf{g}(\mathbf{p})$ é a matriz jacobiana de dimensão $(L-1)(C-1) \times (L-1)(C-1)$ dada por

$$\boldsymbol{\Delta}\mathbf{g}(\mathbf{p}) = [\partial\mathbf{g}/\partial p_{11}, \ldots, \partial\mathbf{g}/\partial p_{1\,C-1}, \ldots, \partial\mathbf{g}/\partial p_{L-1\,1}, \ldots, \partial\mathbf{g}/\partial p_{L-1\,C-1}]$$

e $\hat{\mathbf{V}}_p$ é uma estimativa consistente da matriz de covariância de $\hat{\mathbf{p}}$.

é possível ainda introduzir, no caso de se ter o número m de unidades primárias pequeno, correção na estatística de Wald, utilizando as propostas alternativas de estatísticas F-corrigidas, como em (**??**) e (**??**), com $(L-1)(C-1)$ no lugar de $J-1$, obtendo-se

$$F_{1.p} = \frac{f - (L-1)(C-1) - 1}{f(L-1)(C-1)} X^2_W(I) ,$$

que tem distribuição assintótica **F** com $(L-1)(C-1)$ e $f-(L-1)(C-1)-1$ graus de liberdade e

$$F_{2.p} = \frac{X_W^2(I)}{(L-1)(C-1)},$$

que tem distribuição assintótica **F** com $(L-1)(C-1)$ e f graus de liberdade.

8.3.4 Estatística de Pearson com Ajuste de Rao-Scott

Na presença de efeitos de plano amostral importantes, as estatísticas clássicas de teste precisam ser ajustadas para terem a mesma distribuição assintótica de referência que a obtida para o caso de amostragem aleatória simples.

A estatística de teste de independência $X_P^2(I)$ de Pearson para a tabela $L \times C$ é dada por

$$X_P^2(I) = n \sum_{l=1}^{L} \sum_{c=1}^{C} \frac{\left(\hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c}\right)^2}{\hat{p}_{l+}\hat{p}_{+c}}.$$

Esta estatística pode ser escrita em forma matricial como

$$X_P^2(I) = n\, \hat{\mathbf{f}}'\, \widehat{\mathbf{P}}_{0\mathbf{f}}\, \hat{\mathbf{f}}, \tag{8.12}$$

onde

$$\widehat{\mathbf{P}}_{0\mathbf{f}} = \boldsymbol{\Delta}\mathbf{g}(\hat{\mathbf{p}})\, \hat{\mathbf{P}}_0 \boldsymbol{\Delta}\mathbf{g}(\hat{\mathbf{p}})', \tag{8.13}$$

$$\hat{\mathbf{P}}_0 = diag(\hat{\mathbf{p}}_0) - \hat{\mathbf{p}}_0 \hat{\mathbf{p}}_0^{'},$$

$\widehat{\mathbf{P}}_0/n$ estima a matriz $(L-1)(C-1) \times (L-1)(C-1)$ de covariância multinomial de $\hat{\mathbf{p}}$ sob a hipótese nula, $\hat{\mathbf{p}}_0$ é o vetor com componentes $\hat{p}_{l+}\ \hat{p}_{+c}$, e $diag(\hat{\mathbf{p}}_0)$ representa a matriz diagonal com elementos $\hat{p}_{l+}\ \hat{p}_{+c}$ na diagonal.

Observemos que a forma de $X_P^2(I)$ como expressa em (8.12) é semelhante à da estatística de Wald dada em (8.10), a diferença sendo a estimativa da matriz de covariância de $\hat{\mathbf{f}}$ usada em cada uma dessas estatísticas.

Como nos testes de qualidade de ajuste e de homogeneidade no caso de plano amostral complexo, podemos introduzir correções simples na estatística de Pearson em (8.12) para obter estatísticas de teste com distribuições assintóticas conhecidas.

Inicialmente, vamos considerar ajustes baseados nos efeitos univariados de plano amostral estimados, $\hat{d}_{lc}$, das estimativas das proporções nas celas $\hat{p}_{lc}$. O ajuste mais simples é feito dividindo-se o valor da estatística X_P^2 de Pearson pela média $\hat{d}_{\cdot}$ dos efeitos univariados de plano amostral:

$$X_P^2\left(I; \hat{d}_{\cdot}\right) = X_P^2(I)/\hat{d}_{\cdot},$$

onde $\hat{d}_{\cdot} = \sum_{c=1}^{C}\sum_{l=1}^{L} \hat{d}_{lc}/(LC)$ é um estimador da média dos efeitos univariados de plano amostral desconhecidos.

Estimamos os efeitos do plano amostral por $\hat{d}_{lc} = \hat{V}_p(\hat{p}_{lc}) / (\hat{p}_{lc}(1-\hat{p}_{lc})/n)$, onde $\hat{V}_p(\hat{p}_{lc})$ é a estimativa da variância de aleatorização do estimador de proporção $\hat{p}_{lc}$. Este ajustamento requer que estejam disponíveis as estimativas dos efeitos de plano amostral dos estimadores das proporções nas $L \times C$ celas da tabela.

A seguir vamos apresentar as correções de primeira e de segunda ordem de Rao-Scott para a estatística $X_P^2(I)$ de Pearson para o teste de independência. Estas correções baseiam-se nos autovalores da matriz estimada de efeito multivariado de plano amostral, dada por

$$\hat{\mathbf{\Delta}} = n\, \hat{\mathbf{P}}_{0\mathbf{f}}^{-1}\, \hat{\mathbf{V}}_{\mathbf{f}}, \tag{8.14}$$

onde $\hat{\mathbf{V}}_{\mathbf{f}}$ foi definido em (8.11) e $\hat{\mathbf{P}}_{0\mathbf{f}}$ definido em (8.13).

O ajuste de Rao-Scott de primeira ordem para $X_P^2(I)$ é dado por

$$X_P^2\left(I;\hat{\delta}_{\cdot}\right) = X_P^2(I)/\hat{\delta}_{\cdot}, \tag{8.15}$$

onde $\hat{\delta}_{\cdot}$ é um estimador da média $\bar{\delta}$ dos autovalores desconhecidos da matriz $\mathbf{\Delta}$ de efeitos multivariados de plano amostral.

Podemos estimar a média dos efeitos generalizados, usando os efeitos univariados nas celas e nas marginais da tabela, por

$$\begin{aligned}\hat{\delta}_{\cdot} &= \frac{1}{(L-1)(C-1)} \sum_{l=1}^{L}\sum_{c=1}^{C} \frac{\hat{p}_{lc}(1-\hat{p}_{lc})}{\hat{p}_{l+}\hat{p}_{+c}}\hat{d}_{lc} \\ &\quad - \sum_{l=1}^{L}(1-\hat{p}_{l+})\,\hat{d}_{l+} - \sum_{c=1}^{C}(1-\hat{p}_{+c})\,\hat{d}_{+c},\end{aligned}$$

sem precisar calcular a matriz de efeitos multivariados de plano amostral. A distribuição assintótica de $X_P^2\left(I;\hat{\delta}_{\cdot}\right)$, sob H_0, é qui-quadrado com $(L-1)\times(C-1)$ graus de liberdade.

O ajuste de Rao-Scott de segunda ordem é definido por

$$X_P^2\left(I;\hat{\delta}_{\cdot};\hat{a}^2\right) = X_P^2(I)/\left(\hat{\delta}_{\cdot}\left(1+\hat{a}^2\right)\right),$$

onde $\hat{\delta}_{\cdot}$ é um estimador da média dos autovalores de $\hat{\mathbf{\Delta}}$, dado por

$$\hat{\delta}_{\cdot} = \frac{tr\left(\hat{\mathbf{\Delta}}\right)}{(L-1)(C-1)}$$

e $\hat{a}^2$ é um estimador do quadrado do coeficiente de variação dos autovalores desconhecidos de $\mathbf{\Delta}$, δ_k, $k = 1,\ldots,(L-1)(C-1)$, dado por

$$\hat{a}^2 = \sum_{k=1}^{(L-1)(C-1)} \hat{\delta}_k^2 / \left((L-1)(C-1)\,\hat{\delta}_{\cdot}^2\right) - 1.$$

Um estimador da soma dos quadrados dos autovalores é

$$\sum_{k=1}^{(L-1)(C-1)} \hat{\delta}_k^2 = tr\left(\hat{\mathbf{\Delta}}^2\right).$$

A estatística $X_P^2\left(I;\hat{\delta}_{\cdot};\hat{a}^2\right)$ é assintoticamente qui-quadrado com graus de liberdade com ajuste de Satterthwaite $gl_S = (L-1)(C-1)/\left(1+\hat{a}^2\right)$.

Em situações instáveis, pode ser necessário fazer uma correção F ao ajuste de primeira ordem de Rao-Scott (8.15). A estatística F-corrigida é definida por

$$FX_P^2\left(\hat{\delta}_{\cdot}\right) = X_P^2\left(\hat{\delta}_{\cdot}\right)/(L-1)(C-1) \quad . \tag{8.16}$$

A estatística (8.16) tem distribuição de referência F com $(L-1)\times(C-1)$ e f graus de liberdade.

Tabela 8.4: Frequências amostrais por celas na PNAD 90

	Renda Mensal			
Sexo	1	2	3	**Total**
1	476	2.527	1.273	4.276
2	539	1.270	422	2.231
Total	1.015	3.797	1.695	6.507

Tabela 8.5: Proporções nas linhas, desvios padrões e EPAs

	Renda Mensal			
Sexo	1	2	3	**Total**
1	0,111 57,269 1,420	0,591 102,576 1,861	0,298 111,213 2,527	1,00
2	0,240 125,026 1,909	0,570 119,375 1,297	0,190 111,410 1,800	1,00
Amostra completa	0,155 68,977 2,358	0,584 82,001 1,800	0,261 96,1300 3,117	1,00

Exemplo 8.1 *Correções de EPA médio das estatísticas* $X_P^2(I)$ *e* $X_P^2(H)$.

Considerando os dados do Exemplo **??** do Capítulo 6, vamos testar a hipótese de independência entre as variáveis Sexo (sx) e Rendimento médio mensal (re). Vamos fazer também um teste de homogeneidade, para comparar as distribuições de renda para os dois sexos.

A variável `sx` tem dois níveis: sx(1)-Homens, sx(2)- Mulheres e a variável `re` tem três níveis: re(1)- Menos de salário mínimo, re(2) - de 1 a 5 salário mínimos e re(3)- mais de 5 salários mínimos. A Tabela 8.4 apresenta as frequências nas celas para a amostra pesquisada.

No teste de homogeneidade das distribuições de renda, consideramos fixadas as marginais 4.276 e 2.231 da variável Sexo na tabela de frequências amostrais. Usando o programa Stata, calculamos as estimativas das proporções nas linhas da tabela. Nestas estimativas são considerados os pesos das unidades da amostra e o plano amostral utilizado na pesquisa (PNAD 90), conforme descrito no Exemplo **??**ex:pnad90) do Capítulo 6.

Vamos considerar o teste de homogeneidade entre as variáveis Sexo e Renda e calcular o efeito de plano amostral médio das estimativas das proporções nas celas da tabela. A Tabela 8.5 contém, em cada cela, as estimativas: da proporção na linha, do desvio-padrão da estimativa da proporção na linha ($\times 10.000$), e do efeito de plano amostral da estimativa de proporção na linha.

```
marg_re_pop <-as.data.frame(svymean(~re,pnad.des, deff=TRUE ))
knitr::kable(marg_re_pop ,booktabs=TRUE, digits=3,
  caption="Proporções nas linha, desvios padrões e EPAs de `re` na população")
```

Vamos calcular, a título de ilustração, uma das celas de tabela de efeitos de plano amostral, digamos a cela (1,1). A estimativa da variância do estimador da proporção de linha nesta cela é $(0,0057269)^2$. Sob amostragem aleatória simples com reposição, a estimativa da variância do estimador de proporção de linha

Tabela 8.6: Proporções nas celas, desvios padrões e EPAs

Sexo	Renda Mensal			Total
	1	2	3	
1	0,073 38,343 1,414	0,388 80,435 1,772	0,196 71,772 2,128	0,657 55,814 0,899
2	0,082 44,401 1,695	0,195 51,582 1,101	0,065 40,219 1,729	0,343 55,814 0,899
Total	0,155 68,977 2,358	0,584 82,001 1,800	0,261 96,130 3,117	1,000

na cela é: $0,111\,(1-0,111)\,/4.276$. A estimativa do efeito de plano amostral do estimador de proporção na cela é portanto igual a

$$\frac{(0,0057269)^2}{0,111\,(1-0,111)\,/4.276} \cong 1,420 \ .$$

A estimativa do efeito médio de plano amostral para corrigir a estatística $X_P^2\,(H)$ é $\hat{d}_{\cdot} = 1,802$, calculada tomando a média dos EPAs das celas correspondentes aos níveis 1 e 2 da variável `sx`.

Vamos agora considerar o teste de independência entre as variáveis Sexo e Renda e calcular o efeito de plano amostral médio das estimativas das proporções nas celas da tabela. A Tabela 8.6 contém, em cada cela, as estimativas: da proporção na cela, do desvio-padrão da estimativa da proporção na cela ($\times 10.000$), e do efeito de plano amostral da estimativa de proporção na cela.

Tabela de proporções de `sx` para a população inteira:

```
marg_sx_pop <-data.frame(svymean(~sx,pnad.des, deff=TRUE ))
marg_sx_pop <- transform(marg_sx_pop, SE = 10000*SE)
knitr::kable(marg_sx_pop, digits=3,
  caption="Proporções nas linha, desvios padrões e EPAs de sx na população")
```

Vamos calcular, a título de ilustração, o efeito de plano amostral na cela (1,1) da Tabela 8.6. A estimativa da variância do estimador de proporção nesta cela é $(0,0038343)^2$. Sob amostragem aleatória simples com reposição, a estimativa da variância do estimador de proporção na cela é: $0,073 \times (1-0,073)\,/6.507$. A estimativa do efeito de plano amostral do estimador de proporção na cela é

$$\frac{(0,0038343)^2}{0,073\,(1-0,073)\,/6.507} \cong 1,414 \ .$$

Portanto, a estimativa do efeito médio de plano amostral requerida para corrigir a estatística $X_P^2\,(I)$ é $\hat{d}_{\cdot} = 1,640$, calculada tomando a média dos EPAs das celas correspondentes aos níveis 1 e 2 da variável `sx`.

Calculando as estatísticas $X_P^2\,(I)$ e $X_P^2\,(H)$ para os testes clássicos de independência e homogeneidade a partir da Tabela 8.6, obtemos os valores $X_P^2\,(I) = X_P^2\,(H) = 227,025$, com distribuição de referência $\chi^2\,(2)$, resultado que indica rejeição da hipótese de independência entre `sx` e `re`, bem como da hipótese de igualdade de distribuição de renda para os dois sexos a partir do teste de homogeneidade. O valor comum das estatísticas $X_P^2\,(I)$ e $X_P^2\,(H)$ foi calculado sem considerar os pesos e o plano amostral. Considerando estes últimos,

mediante a correção de EPA médio das estatísticas clássicas, obtemos os valores $X^2_P\left(I;\hat{d}_{\cdot}\right) = 137,117$ e $X^2_P\left(H;\hat{d}_{\cdot}\right) = 124,742$, que também indicam a rejeição das hipóteses de independência e de homogeneidade.

Vale ressaltar que apesar de todos os testes mencionados indicarem forte rejeição das hipóteses de independência e de homogeneidade, os valores das estatísticas de teste $137,117$ e $124,742$, calculados considerando os pesos e plano amostral, são bem menores que o valor $227,025$ obtido para o caso de amostra IID. Sob a hipótese nula, a distribuição de referência de todas essas estatísticas de teste é $\chi^2(2)$, mostrando novamente que a estatística de teste calculada sob a hipótese de amostra IID tem maior tendência a rejeitar a hipótese nula.

A partir da Tabela 8.6, examinando as estimativas das proporções nas celas da tabela para cada sexo, observamos uma ordenação estocástica das distribuições de renda para os dois sexos, com proporções maiores em valores mais altos para o nível 1 da variável sexo, que é o sexo masculino.

8.4 Laboratório de R

Vamos reproduzir alguns resultados usando dados da PNAD descritos na Seção ???

Exemplo 8.2 *Estimativas de medidas descritivas em tabelas*

```
library(survey)
```

```
## Loading required package: grid
```

```
## Loading required package: Matrix
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survey'
```

```
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
pnadrj90 <- readRDS("~\\GitHub\\adac\\data\\pnadrj90.rds")
names(pnadrj90)
```

```
##  [1] "stra"     "psu"      "pesopes"  "informal" "sx"       "id"
##  [7] "ae"       "ht"       "re"       "um"
```

Transformação em fatores:

```
unlist(lapply(pnadrj90, mode))
```

```
##      stra       psu   pesopes  informal        sx        id        ae
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        ht        re        um
## "numeric" "numeric" "numeric"
```

```
pnadrj90<-transform(pnadrj90,sx=factor(sx),id=factor(id),ae=factor(ae),ht=factor(ht),re=factor(re))
```

Definição do objeto de desenho:

```
pnad.des<-svydesign(id=~psu,strata=~stra,weights=~pesopes,data=pnadrj90,nest=TRUE)
```

Estimativas de proporções:

```r
svymean(~sx,pnad.des)          #estimativa de proporção para sx
```

```
##        mean     SE
## sx1 0.65708 0.0056
## sx2 0.34292 0.0056
```

```r
svymean(~re,pnad.des)          #estimativa de proporção para re
```

```
##        mean     SE
## re1 0.15546 0.0069
## re2 0.58356 0.0082
## re3 0.26098 0.0096
```

```r
svymean(~ae,pnad.des)          #estimativa de proporção para ae
```

```
##        mean     SE
## ae1 0.31304 0.0095
## ae2 0.31972 0.0071
## ae3 0.36725 0.0105
```

```r
ht.mean<-svymean(~ht,pnad.des)
```

Exemplos de funções extratoras e atributos

```r
coef(ht.mean)                          #estimativas das proporções
```

```
##       ht1       ht2       ht3
## 0.2103714 0.6148881 0.1747405
```

```r
attributes(ht.mean)                          #ver atributos
```

```
## $names
## [1] "ht1" "ht2" "ht3"
##
## $var
##                ht1           ht2           ht3
## ht1  3.666206e-05 -3.322546e-05 -3.436592e-06
## ht2 -3.322546e-05  6.758652e-05 -3.436106e-05
## ht3 -3.436592e-06 -3.436106e-05  3.779765e-05
##
## $statistic
## [1] "mean"
##
## $class
## [1] "svystat"
```

```r
vcov(ht.mean)                          #estimativas de variâncias e covariâncias
```

```
##                ht1           ht2           ht3
## ht1  3.666206e-05 -3.322546e-05 -3.436592e-06
## ht2 -3.322546e-05  6.758652e-05 -3.436106e-05
## ht3 -3.436592e-06 -3.436106e-05  3.779765e-05
```

```r
attr(ht.mean, "var")
```

```
##                ht1           ht2           ht3
## ht1  3.666206e-05 -3.322546e-05 -3.436592e-06
## ht2 -3.322546e-05  6.758652e-05 -3.436106e-05
## ht3 -3.436592e-06 -3.436106e-05  3.779765e-05
```

Podemos obter estimativas de proporções nas classes de renda por domínios definidos pela variável **sx**:

```
library(xtable)
print(xtable(svyby(~re,~sx,pnad.des,svymean,keep.var=TRUE),caption="Proporções por sexo"),type="html", s
```

As proporções estimadas nas classes de renda e a tabela cruzada das variáveis sexo e renda são obtidas a seguir:

```
svymean(~re,pnad.des,deff=T)
```

```
##          mean        SE   DEff
## re1 0.1554555 0.0068977 2.3611
## re2 0.5835630 0.0082001 1.8027
## re3 0.2609815 0.0096130 3.1216
```

```
round(svytable(~sx+re,pnad.des,Ntotal=1),digits=3)
```

```
##    re
## sx     1     2     3
##   1 0.073 0.388 0.196
##   2 0.082 0.195 0.065
```

```
svyby(~re,~sx,pnad.des,svymean,keep.var=T)
```

```
##   sx       re1       re2       re3      se.re1     se.re2     se.re3
## 1  1 0.1110831 0.5908215 0.2980955 0.005726888 0.01025759 0.01112131
## 2  2 0.2404788 0.5696548 0.1898663 0.012502636 0.01193753 0.01114102
```

```
svymean(~I((sx==1&re==1)*1),pnad.des,deff=T)
```

```
##                                  mean        SE   DEff
## I((sx == 1 & re == 1) * 1) 0.0729904 0.0038343 1.4156
```

```
#proporções nas celas
svytable(~sx+re,pnad.des,Ntotal=1)
```

```
##    re
## sx          1          2          3
##   1 0.07299044 0.38821684 0.19587250
##   2 0.08246505 0.19534616 0.06510900
```

```
# porcentagens nas celas
svytable(~sx+re,pnad.des,Ntotal=100)
```

```
##    re
## sx         1         2         3
##   1  7.299044 38.821684 19.587250
##   2  8.246505 19.534616  6.510900
```

```
# produz se e deff
sx.re_mean<-svymean(~interaction(sx,re),pnad.des,deff=T)
```

Resultados dos testes obtidos pela library **survey** (Lumley, 2016) precisam ser identificados com as fórmulas do texto:

```
# teste Chi-quadrado de Pearson com ajuste de Rao-Scott
svychisq( ~ sx+re , pnad.des, statistic = "F", na.rm=TRUE)
```

```
##
##  Pearson's X^2: Rao & Scott adjustment
##
```

```
## data:  svychisq(~sx + re, pnad.des, statistic = "F", na.rm = TRUE)
## F = 108.34, ndf = 1.9779, ddf = 1275.8000, p-value < 2.2e-16
```

```
# teste Chi-quadrado de Pearson com ajuste de Rao-Scott
svychisq( ~ sx+re , pnad.des, statistic = "Chisq", na.rm=TRUE)
```

```
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~sx + re, pnad.des, statistic = "Chisq", na.rm = TRUE)
## X-squared = 224.85, df = 2, p-value < 2.2e-16
```

```
# teste de Wald baseado no desenho amostral
svychisq( ~ sx+re , pnad.des, statistic = "Wald", na.rm=TRUE)
```

```
##
##  Design-based Wald test of association
##
## data:  svychisq(~sx + re, pnad.des, statistic = "Wald", na.rm = TRUE)
## F = 68.41, ndf = 2, ddf = 645, p-value < 2.2e-16
```

```
# teste de Wald com ajuste
svychisq( ~ sx+re , pnad.des, statistic = "adjWald", na.rm=TRUE)
```

```
##
##  Design-based Wald test of association
##
## data:  svychisq(~sx + re, pnad.des, statistic = "adjWald", na.rm = TRUE)
## F = 68.304, ndf = 2, ddf = 644, p-value < 2.2e-16
```

```
# teste Chi-quadrado de Pearson: distribuição assintótica exata
svychisq( ~ sx+re , pnad.des, statistic = "lincom", na.rm=TRUE)
```

```
## Warning in pFsum(pearson$statistic, rep(1, ncol(Delta)), eigen(Delta,
## only.values = TRUE)$values, : Package 'CompQuadForm' not found, using
## saddlepoint approximation
```

```
##
##  Pearson's X^2: asymptotic exact distribution
##
## data:  svychisq(~sx + re, pnad.des, statistic = "lincom", na.rm = TRUE)
## X-squared = 224.85, p-value < 2.2e-16
```

```
# teste Chi-quadrado de Pearson: aproximação de ponto de cela
svychisq( ~ sx+re , pnad.des, statistic = "saddlepoint", na.rm=TRUE)
```

```
##
##  Pearson's X^2: saddlepoint approximation
##
## data:  svychisq(~sx + re, pnad.des, statistic = "saddlepoint", na.rm = TRUE)
## X-squared = 224.85, p-value < 2.2e-16
```

Capítulo 9

Estimação de densidades

Capítulo 10

Modelos Hierárquicos

Capítulo 11

Não-Resposta

Capítulo 12

Diagnóstico de ajuste de modelo

Capítulo 13

Agregação vs. Desagregação

Capítulo 14

Pacotes para Analisar Dados Amostrais

Capítulo 15

Placeholder

Referências Bibliográficas

Lumley, T. (2016). survey: analysis of complex survey samples. R package version 3.31-5.