Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-09-06

Contents

Prefácio **9**
Agradecimentos . . . 9

1 Introdução **11**
1.1 Motivação . . . 12
1.2 Objetivos do Livro . . . 12
1.3 Laboratório de R do Capítulo 1. . . . 12
1.4 total SE DEff . . . 12
1.5 analf1 1174220 127982 2.0543 . . . 12
1.6 total SE DEff . . . 12
1.7 analf2 4792344 318877 3.3237 . . . 12
1.8 Ratio estimator: svyratio.survey.design2(~analf1, ~faixa1, ppv.se.des) . . . 12
1.9 Ratios= . . . 12
1.10 faixa1 . . . 12
1.11 analf1 0.118689 . . . 12
1.12 SEs= . . . 12
1.13 faixa1 . . . 12
1.14 analf1 0.01178896 . . . 12
1.15 Ratio estimator: svyratio.survey.design2(~analf2, ~faixa2, ppv.se.des) . . . 12
1.16 Ratios= . . . 12
1.17 faixa2 . . . 12
1.18 analf2 0.1086871 . . . 12
1.19 SEs= . . . 12
1.20 faixa2 . . . 12
1.21 analf2 0.006732254 . . . 12
1.22 regiao analf1 se DEff.analf1 . . . 12
1.23 1 1 3512866 352619.5 9.660561 . . . 12
1.24 2 2 1174220 127982.2 2.054345 . . . 12
1.25 analf1 . . . 12
1.26 analf1 2.054049 . . . 12
1.27 analf2 . . . 12
1.28 analf2 3.32324 . . . 12
1.29 total SE DEff . . . 12
1.30 [1,] 1174220 127982 2.6426 . . . 12
1.31 total SE DEff . . . 12
1.32 [1,] 4792344 318877 4.1667 . . . 12
1.33 total SE DEff . . . 12
1.34 I(ifelse(regiao == 2, analf1, 0)) 1174220 127982 2.6426 . . . 12
1.35 total SE DEff . . . 12
1.36 I(ifelse(regiao == 2, analf2, 0)) 4792344 318877 4.1667 . . . 12
1.37 Taxa Erro_Padrao CV . . . 12
1.38 1 11.80303 0.1174791 0.9953299 . . . 12

1.39 [1] 60202 1019 . . . 12
1.40 diag_dep . . . 12
1.41 7.6 . . . 12
1.42 diag_dep . . . 12
1.43 diag_dep 0.2 . . . 12
1.44 [1] 7.2 8.1 . . . 12
1.45 diag_dep . . . 12
1.46 7.6 7.2 8.1 . . . 12

2 Referencial para Inferência **13**
2.1 Modelagem - Primeiras Idéias . . . 13
2.2 Fontes de Variação . . . 13
2.3 Modelos de Superpopulação . . . 13
2.4 Planejamento Amostral . . . 13
2.5 Planos Amostrais Informativos e Ignoráveis . . . 13

3 Estimação Baseada no Plano Amostral **15**
3.1 Estimação de Totais . . . 16
3.2 Por que Estimar Variâncias . . . 16
3.3 Linearização de Taylor para Estimar variâncias . . . 16
3.4 Método do Conglomerado Primário . . . 16
3.5 Métodos de Replicação . . . 16
3.6 Laboratório de R . . . 16
3.7 faixa . . . 16
3.8 0.01178896 . . . 16
3.9 Ratio estimator: svyratio.survey.design2(~analf.faixa, ~faixa, ppv.se.des) . . . 16
3.10 Ratios= . . . 16
3.11 faixa . . . 16
3.12 analf.faixa 0.118689 . . . 16
3.13 SEs= . . . 16
3.14 faixa . . . 16
3.15 analf.faixa 0.01178896 . . . 16
3.16 Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv.se.des.jkn) . . . 16
3.17 Ratios= . . . 16
3.18 faixa . . . 16
3.19 analf.faixa 0.118689 . . . 16
3.20 SEs= . . . 16
3.21 [,1] . . . 16
3.22 [1,] 0.01181434 . . . 16
3.23 Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv.se.des.boot) . . . 16
3.24 Ratios= . . . 16
3.25 faixa . . . 16
3.26 analf.faixa 0.118689 . . . 16
3.27 SEs= . . . 16
3.28 [,1] . . . 16
3.29 [1,] 0.01256654 . . . 16
3.30 [1] ``svyrep.design'' . . . 16
3.31 [1] ``repweights'' ``pweights'' ``type'' . . . 16
3.32 [4] ``rho'' ``scale'' ``rscales'' . . . 16
3.33 [7] ``call'' ``combined.weights'' ``selfrep'' . . . 16
3.34 [10] ``mse'' ``variables'' ``degf'' . . . 16
3.35 [1] 8903 . . . 16
3.36 [1] 276 . . . 16
3.37 num [1:8903, 1:276] 0 0 1.06 1.06 1.06 … . . . 16

3.38 [1] 0.01181 . . . 16
3.39 theta SE . . . 16
3.40 [1,] 0.11869 0.0118 . . . 16
3.41 theta SE . . . 16
3.42 [1,] 0.50623 0.0494 . . . 16

4 Efeitos do Plano Amostral **17**
4.1 Introdução . . . 18
4.2 Efeito do Plano Amostral (EPA) de Kish . . . 18
4.3 Efeito do Plano Amostral Ampliado . . . 18
4.4 Intervalos de Confiança e Testes de Hipóteses . . . 18
4.5 Efeitos Multivariados de Plano Amostral . . . 18
4.6 Laboratório de R . . . 18
4.7 Warning: package `survey' was built under R version 3.4.1 . . . 18
4.8 Carregando pacotes exigidos: grid . . . 18
4.9 Carregando pacotes exigidos: methods . . . 18
4.10 Carregando pacotes exigidos: Matrix . . . 18
4.11 Carregando pacotes exigidos: survival . . . 18
4.12 . . . 18
4.13 Attaching package: `survey' . . . 18
4.14 The following object is masked from `package:graphics': . . . 18
4.15 . . . 18
4.16 dotchart . . . 18
4.17 [1] 0.140409 . . . 18
4.18 [1] 0.3963556 . . . 18

5 Ajuste de Modelos Paramétricos **19**
5.1 Introdução . . . 19
5.2 Método de Máxima Verossimilhança (MV) . . . 19
5.3 Ponderação de Dados Amostrais . . . 19
5.4 Método de Máxima Pseudo-Verossimilhança . . . 19
5.5 Robustez do Procedimento MPV . . . 19
5.6 Desvantagens da Inferência de Aleatorização . . . 19
5.7 Laboratório de R . . . 19

6 Modelos de Regressão **21**
6.1 Modelo de Regressão Linear Normal . . . 22
6.2 Modelo de Regressão Logística . . . 22
6.3 Warning in 1.96 * sqrt(var_raz112) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated. . . . 22
6.4 Use c() or as.vector() instead. . . . 22
6.5 Warning in 1.96 * sqrt(var_raz123) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated. . . . 22
6.6 Use c() or as.vector() instead. . . . 22
6.7 Warning in 1.96 * sqrt(var_raz212) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated. . . . 22
6.8 Use c() or as.vector() instead. . . . 22
6.9 Warning in 1.96 * sqrt(var_raz223) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated. . . . 22
6.10 Use c() or as.vector() instead. . . . 22
6.11 Warning in 1.96 * sqrt(var_raz312) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated. . . . 22
6.12 Use c() or as.vector() instead. . . . 22

6.13 Warning in 1.96 * sqrt(var_raz323) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated. . . . 22
6.14 Use c() or as.vector() instead. . . . 22
6.15 Teste de Hipóteses . . . 22
6.16 Laboratório de R . . . 22
6.17 [1] ``stra'' ``psu'' ``pesopes'' ``informal'' ``sx'' ``id'' . . . 22
6.18 [7] ``ae'' ``ht'' ``re'' ``um'' . . . 22
6.19 stra psu pesopes informal sx id ae . . . 22
6.20 ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' . . . 22
6.21 ht re um . . . 22
6.22 ``numeric'' ``numeric'' ``numeric'' . . . 22
6.23 Wald test for ht:re . . . 22
6.24 in svyglm(formula = informal ~ sx + ae + ht + id + re + sx * id + . . . 22
6.25 sx * ht + ae * ht + ht * id + ht * re, design = pnad.des, . . . 22
6.26 family = binomial) . . . 22
6.27 F = 6.742662 on 4 and 616 df: p= 2.58e-05 . . . 22

7 Testes de Qualidade de Ajuste **23**
7.1 Introdução . . . 23
7.2 Teste para uma Proporção . . . 23
7.3 Teste para Várias Proporções . . . 28
7.4 Laboratório de R . . . 36

8 Testes em Tabelas de Duas Entradas **39**
8.1 Introdução . . . 40
8.2 Tabelas 2x2 . . . 40
8.3 Tabelas de Duas Entradas (Caso Geral) . . . 40
8.4 Laboratório de R . . . 40
8.5 [1] ``stra'' ``psu'' ``pesopes'' ``informal'' ``sx'' ``id'' . . . 40
8.6 [7] ``ae'' ``ht'' ``re'' ``um'' . . . 40
8.7 stra psu pesopes informal sx id ae . . . 40
8.8 ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' . . . 40
8.9 ht re um . . . 40
8.10 ``numeric'' ``numeric'' ``numeric'' . . . 40
8.11 mean SE . . . 40
8.12 sx1 0.65708 0.0056 . . . 40
8.13 sx2 0.34292 0.0056 . . . 40
8.14 mean SE . . . 40
8.15 re1 0.15546 0.0069 . . . 40
8.16 re2 0.58356 0.0082 . . . 40
8.17 re3 0.26098 0.0096 . . . 40
8.18 mean SE . . . 40
8.19 ae1 0.31304 0.0095 . . . 40
8.20 ae2 0.31972 0.0071 . . . 40
8.21 ae3 0.36725 0.0105 . . . 40
8.22 ht1 ht2 ht3 . . . 40
8.23 0.2103714 0.6148881 0.1747405 . . . 40
8.24 $names . . . 40
8.25 [1] ``ht1'' ``ht2'' ``ht3'' . . . 40
8.26 . . . 40
8.27 $var . . . 40
8.28 ht1 ht2 ht3 . . . 40
8.29 ht1 3.666206e-05 -3.322546e-05 -3.436592e-06 . . . 40
8.30 ht2 -3.322546e-05 6.758652e-05 -3.436106e-05 . . . 40

8.31 ht3 -3.436592e-06 -3.436106e-05 3.779765e-05 . . . 40
8.32 . . . 40
8.33 $statistic . . . 40
8.34 [1] ``mean'' . . . 40
8.35 . . . 40
8.36 $class . . . 40
8.37 [1] ``svystat'' . . . 40
8.38 ht1 ht2 ht3 . . . 40
8.39 ht1 3.666206e-05 -3.322546e-05 -3.436592e-06 . . . 40
8.40 ht2 -3.322546e-05 6.758652e-05 -3.436106e-05 . . . 40
8.41 ht3 -3.436592e-06 -3.436106e-05 3.779765e-05 . . . 40
8.42 ht1 ht2 ht3 . . . 40
8.43 ht1 3.666206e-05 -3.322546e-05 -3.436592e-06 . . . 40
8.44 ht2 -3.322546e-05 6.758652e-05 -3.436106e-05 . . . 40
8.45 ht3 -3.436592e-06 -3.436106e-05 3.779765e-05 . . . 40
8.46 mean SE DEff . . . 40
8.47 re1 0.1554555 0.0068977 2.3611 . . . 40
8.48 re2 0.5835630 0.0082001 1.8027 . . . 40
8.49 re3 0.2609815 0.0096130 3.1216 . . . 40
8.50 re . . . 40
8.51 sx 1 2 3 . . . 40
8.52 1 0.073 0.388 0.196 . . . 40
8.53 2 0.082 0.195 0.065 . . . 40
8.54 sx re1 re2 re3 se.re1 se.re2 se.re3 . . . 40
8.55 1 1 0.1110831 0.5908215 0.2980955 0.005726888 0.01025759 0.01112131 . . . 40
8.56 2 2 0.2404788 0.5696548 0.1898663 0.012502636 0.01193753 0.01114102 . . . 40
8.57 mean SE DEff . . . 40
8.58 I((sx == 1 & re == 1) * 1) 0.0729904 0.0038343 1.4156 . . . 40
8.59 re . . . 40
8.60 sx 1 2 3 . . . 40
8.61 1 0.07299044 0.38821684 0.19587250 . . . 40
8.62 2 0.08246505 0.19534616 0.06510900 . . . 40
8.63 re . . . 40
8.64 sx 1 2 3 . . . 40
8.65 1 7.299044 38.821684 19.587250 . . . 40
8.66 2 8.246505 19.534616 6.510900 . . . 40
8.67 [1] 1.642161 . . . 40
8.68 . . . 40
8.69 Pearson's Chi-squared test . . . 40
8.70 . . . 40
8.71 data: tab.amo . . . 40
8.72 X-squared = 227.03, df = 2, p-value < 2.2e-16 . . . 40

9 Estimação de densidades **41**
9.1 Introdução . . . 41

10 Modelos Hierárquicos **43**
10.1 Introdução . . . 43

11 Não-Resposta **45**
11.1 Introdução . . . 45

12 Diagnóstico de ajuste de modelo **47**
12.1 Introdução . . . 47

13 Agregação vs. Desagregação **49**
13.1 Introdução 49
13.2 Modelagem da Estrutura Populacional 49
13.3 Modelos Hierárquicos 49
13.4 Análise Desagregada: Prós e Contras 49

14 Pacotes para Analisar Dados Amostrais **51**
14.1 Introdução 51
14.2 Pacotes Computacionais 51

Referências **53**

Prefácio

Agradecimentos

Chapter 1

Introdução

1.1 Motivação

1.2 Objetivos do Livro

1.3 Laboratório de R do Capítulo 1.

1.4 total SE DEff

1.5 analf1 1174220 127982 2.0543

1.6 total SE DEff

1.7 analf2 4792344 318877 3.3237

1.8 Ratio estimator: svyratio.survey.design2(˜analf1, ˜faixa1, ppv.se.des)

1.9 Ratios=

1.10 faixa1

1.11 analf1 0.118689

1.12 SEs=

1.13 faixa1

1.14 analf1 0.01178896

1.15 Ratio estimator: svyratio.survey.design2(˜analf2, ˜faixa2, ppv.se.des)

Chapter 2

Referencial para Inferência

2.1 Modelagem - Primeiras Idéias

2.1.1 Abordagem 1 - Modelagem Clássica

2.1.2 Abordagem 2 - Amostragem Probabilística

2.1.3 Discussão das Abordagens 1 e 2

2.1.4 Abordagem 3 - Modelagem de Superpopulação

2.2 Fontes de Variação

2.3 Modelos de Superpopulação

2.4 Planejamento Amostral

2.5 Planos Amostrais Informativos e Ignoráveis

Chapter 3

Estimação Baseada no Plano Amostral

3.1 Estimação de Totais

3.2 Por que Estimar Variâncias

3.3 Linearização de Taylor para Estimar variâncias

3.4 Método do Conglomerado Primário

3.5 Métodos de Replicação

3.6 Laboratório de R

3.7 faixa

3.8 0.01178896

3.9 Ratio estimator: svyratio.survey.design2(~analf.faixa, ~faixa, ppv.se.des)

3.10 Ratios=

3.11 faixa

3.12 analf.faixa 0.118689

3.13 SEs=

3.14 faixa

Chapter 4

Efeitos do Plano Amostral

4.1 Introdução

4.2 Efeito do Plano Amostral (EPA) de Kish

4.3 Efeito do Plano Amostral Ampliado

4.4 Intervalos de Confiança e Testes de Hipóteses

4.5 Efeitos Multivariados de Plano Amostral

4.6 Laboratório de R

4.7 Warning: package `survey' was built under R version 3.4.1

4.8 Carregando pacotes exigidos: grid

4.9 Carregando pacotes exigidos: methods

4.10 Carregando pacotes exigidos: Matrix

4.11 Carregando pacotes exigidos: survival

4.12

4.13 Attaching package: `survey'

4.14 The following object is masked from `package:graphics':

4.15

4.16 dotchart

Chapter 5

Ajuste de Modelos Paramétricos

5.1 Introdução

5.2 Método de Máxima Verossimilhança (MV)

5.3 Ponderação de Dados Amostrais

5.4 Método de Máxima Pseudo-Verossimilhança

5.5 Robustez do Procedimento MPV

5.6 Desvantagens da Inferência de Aleatorização

5.7 Laboratório de R

Chapter 6

Modelos de Regressão

6.1 Modelo de Regressão Linear Normal

6.1.1 Especificação do Modelo

6.1.2 Pseudo-parâmetros do Modelo

6.1.3 Estimadores de MPV dos Parâmetros do Modelo

6.1.4 Estimação da Variância de Estimadores de MPV

6.2 Modelo de Regressão Logística

6.3 Warning in 1.96 * sqrt(var_raz112) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated.

6.4 Use c() or as.vector() instead.

6.5 Warning in 1.96 * sqrt(var_raz123) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated.

6.6 Use c() or as.vector() instead.

6.7 Warning in 1.96 * sqrt(var_raz212) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated.

6.8 Use c() or as.vector() instead.

6.9 Warning in 1.96 * sqrt(var_raz223) * c(-1, 1): Recycling array of length 1 in array-vector arithmetic is deprecated.

6.10 Use c() or as.vector() instead.

6.11 Warning in 1.96 * sqrt(var_raz312) * c(-1, 1): Recycling array

Chapter 7

Testes de Qualidade de Ajuste

7.1 Introdução

Tabelas de distribuições de frequências ocorrem comumente na análise de dados de pesquisas complexas. Tais tabelas são formadas pela classificação e cálculo de frequências dos dados da amostra disponível segundo níveis de uma variável categórica - tabelas de uma entrada - ou segundo celas de uma classificação cruzada de duas (ou mais) variáveis categóricas - tabelas de duas (ou mais) entradas. Neste capítulo concentraremos a atenção em tabelas de uma entrada, ou equivalentemente nas frequências absolutas e relativas (ou proporções) correspondentes.

Em muitos casos, o objetivo da análise é testar hipóteses de bondade de ajuste de modelos para descrever essas distribuições de frequências. Sob a hipótese de observações IID (distribuição Multinomial) ou equivalentemente, de amostragem aleatória simples, inferências válidas para testar tais hipóteses podem ser baseadas na estatística padrão de teste qui-quadrado de Pearson. Tais testes podem ser facilmente executados usando procedimentos prontos em pacotes estatísticos padrões tais como o SAS, S-Plus, SPSS, GLIM e outros.

No caso de planos amostrais complexos, entretanto, os procedimentos de teste precisam ser ajustados devido aos efeitos de conglomeração, estratificação e/ou pesos desiguais. Neste capítulo examinaremos o impacto do plano amostral sobre as estatísticas de teste usuais notando que, em alguns casos, os valores observados dessas estatísticas de teste podem ser muito grandes, acarretando inferências incorretas, conforme já ilustrado no Exemplo @ref{ex:exebin}. Isto ocorre porque a probabilidade de erros do tipo I (rejeitar a hipótese nula quando esta é verdadeira) é muito maior que o nível nominal de significância α especificado.

Para obter inferências válidas usando amostras complexas podemos introduzir correções na estatística de teste de Pearson, tais como os ajustes de Rao-Scott, ou alternativamente usar outras estatísticas de teste que já incorporem o plano amostral, tais como a estatística de Wald. Os dois enfoques serão ilustrados através de um exemplo introdutório simples de teste de bondade de ajuste. Os resultados discutidos neste capítulo são adequados tanto para uma abordagem de aleatorização, em que os parâmetros se referem à população finita em questão, quanto para uma abordagem baseada em modelos, em que os parâmetros especificam algum modelo de superpopulação.

7.2 Teste para uma Proporção

7.2.1 Correção de Estatísticas Clássicas

No Exemplo @ref{ex:exebin} a estatística de teste Z_{bin}, que foi utilizada para comparar com um valor hipotético pré-fixado a proporção de empregados cobertos por plano de saúde, resultou num teste mais liberal do que o teste que empregou a estatística Z_p, baseada no plano amostral efetivamente adotado. A

causa disto foi o fato de Z_{bin} não considerar o efeito de conglomeração existente. Vamos examinar com mais detalhes o comportamento assintótico da estatística de teste Z_{bin}, construindo a estatística de teste X_P^2 de Pearson para o exemplo correspondente. Para isto, consideremos a Tabela 7.1 contendo a distribuição de frequências, onde n_j e p_{oj} são as frequências (absolutas) observadas na amostra e as proporções hipotéticas nas categorias de interesse, respectivamente.

Table 7.1: Frequências observadas e proporções hipotéticas

Categoria	j	n_j	p_{0j}
Cobertos por planos de saúde	1	840	0.8
Não cobertos	2	160	0.2
Todos empregados	-	1000	1.0

As proporções populacionais desconhecidas nas categorias são $p_j = N_j/N$, onde N é o tamanho total da população de empregados e N_j é o número de elementos da população na categoria j, $j = 1, 2$. Os parâmetros populacionais p_j poderiam também ser considerados como pseudo-parâmetros, se vistos como estimativas de censo para as probabilidades desconhecidas (π_j, digamos) no contexto de um modelo de superpopulação.

A estatística de teste de Pearson para a hipótese simples de bondade de ajuste $H_0 : p_j = p_{0j},\ j = 1, 2$, é dada por

$$X_P^2 = \sum_{j=1}^{2} \left(n_j - n\,p_{0j}\right)^2 / \left(n\,p_{0j}\right) = n \sum_{j=1}^{2} \left(\widehat{p}_j - p_{0j}\right)^2 / p_{0j}, \tag{7.1}$$

onde as proporções $\widehat{p}_j = n_j/n$ são estimativas amostrais usuais das proporções populacionais p_j, para $j = 1, 2$.

```
p0 <- c(.8, .2)
obs<- c(840, 160)
n<- sum(obs)
phat <- obs/n
# Estatística de Pearson
x2p <- sum((obs-n*p0)^2/(n*p0))
x2p
```

```
## [1] 10
```

Como há apenas duas categorias e as proporções devem somar 1, observa-se que $p_2 = 1 - p_1$, $\widehat{p}_2 = 1 - \widehat{p}_1$ e $p_{02} = 1 - p_{01}$. Isto acarreta na equivalência entre as estatísticas Z_{bin} e X_P^2 demonstrada pela relação

$$X_P^2 = n \sum_{j=1}^{2} \left(\widehat{p}_j - p_{0j}\right)^2 / p_{0j} = \frac{\left(\widehat{p} - p_0\right)^2}{p_0\left(1 - p_0\right)/n} = Z_{bin}^2 \tag{7.2}$$

onde $\widehat{p} = \widehat{p}_1$ e $p_0 = p_{01}$ para simplicidade e coerência com a notação do Exemplo @ref{ex:exebin}.

Sob a hipótese de observações IID, a distribuição assintótica da estatística X_P^2 é qui-quadrado (χ^2). Neste caso, em que há apenas duas categorias e uma restrição (soma das proporções igual a 1), a distribuição da estatística X_P^2 em (7.2) tem apenas um grau de liberdade.

Rao e Scott(1981) obtiveram resultados gerais para a distribuição assintótica da estatística de teste X_P^2 de Pearson sob planos amostrais complexos. Com apenas duas celas, a distribuição assintótica da estatística de teste X_P^2 é a distribuição da variável aleatória dW, onde W tem distribuição $\chi^2\,(1)$ (qui-quadrado com um grau de liberdade) e d é o efeito de plano amostral (EPA) da estimativa $\widehat{p}$ da proporção p. O efeito de plano amostral nesse caso é dado por $d = V_p\left(\widehat{p}\right)/V_{bin}\left(\widehat{p}\right)$.

Para uma amostra de empregados selecionada por amostragem aleatória simples, teríamos $d = 1$ pois $V_p(\widehat{p})$ e $V_{bin}(\widehat{p})$ seriam iguais. Neste caso, a estatística X_P^2 de teste seria assintoticamente $\chi^2(1)$. Como a amostra foi efetivamente selecionada por amostragem de conglomerados, devido à correlação intraclasse positiva o efeito de plano amostral d é maior que um, e portanto a distribuição assintótica da estatística de teste X_P^2 não é mais $\chi^2(1)$.

Considerando que o impacto da correlação intraclasse positiva na distribuição assintótica da estatística X_P^2 de Pearson pode levar a inferências incorretas caso se utilize a distribuição assintótica usual, o próximo passo é derivar um procedimento de teste válido. Isto é feito introduzindo uma correção em X_P^2. Para isto, observe que a esperança assintótica de X_P^2 é $E_p\left(X_P^2\right) = d$. Como $E_p\left(X_P^2/d\right) = E\left(\chi^2(1)\right) = 1$, obtemos então a correção simples de Rao-Scott para X_P^2 dividindo o valor observado da estatística de teste pelo efeito do plano amostral d, isto é,

$$X_P^2(d) = X_P^2/d, \tag{7.3}$$

que tem, no caso de duas celas, distribuição assintótica $\chi^2(1)$.

Outra estatística comumente usada para testar a mesma hipótese de bondade de ajuste no caso de proporções é a estatística do teste da Razão de Verossimilhança (RV), dada por

$$X_{RV}^2 = 2n\sum_{j=1}^{2}\widehat{p}_j \log\left(\widehat{p}_j/p_{0j}\right) = 2n\log\left(\frac{\widehat{p}\left(1-\widehat{p}\right)}{p_0\left(1-p_0\right)}\right). \tag{7.4}$$

No caso de amostragem aleatória simples, a estatística X_{RV}^2 é também distribuída assintoticamente como $\chi^2(1)$, quando a hipótese nula é verdadeira. Para planos amostrais complexos, a estatística corrigida correspondente é

$$X_{RV}^2(d) = X_{RV}^2/d \ . \tag{7.5}$$

Vamos calcular os valores das estatísticas de Pearson e de RV, com suas correções de Rao-Scott, para os dados do Exemplo @ref{ex:exebin}. Para as correções, primeiro é preciso calcular o efeito do plano amostral

```
p <- p0[1]
m <- 50
# Efeito do plano amostral
d <- (p*(1-p)/m)/ (p*(1-p)/n)
d
```

```
## [1] 20
```

```
# Estatística de Peason corrigida
x2pd <- x2p/d
x2pd
```

```
## [1] 0.5
```

```
# valor-p do teste
pchisq(x2pd, 1, lower.tail = F)
```

```
## [1] 0.4795001
```

$$\begin{array}{rcccl} d & = & V_p(\widehat{p})/V_{bin}(\widehat{p}) & = & \frac{p(1-p)/m}{p(1-p)/n} \\ & = & \frac{0{,}0032}{0{,}00016} & = & 20 \end{array}$$

onde $m = 50$ é o número de empregados por empresa (tamanho do conglomerado) e $n = 1.000$ é o número de empregados na amostra.

O valor da estatística de teste de Pearson é

$$X_P^2 = \frac{(0,84 - 0,80)^2}{(0,80 \times 0,20)/1.000} = 10$$

com $p$valor $0,0016$. O valor da estatística de teste de Pearson com a correção de Rao-Scott $X_P^2(d)$ é então dado por

$$X_P^2(d) = X_P^2/d = 10/20 = 0,5$$

com $p$valor $0,4795$. Observe que $Z_p^2 = 0,707^2 = 0,5$, e também que $X_P^2(d) = Z_{bin}^2/d = 3,162^2/20 = 0,5$ ou seja, $Z_p^2 = X_P^2(d)$ conforme esperado. Os valores da estatística do teste da Razão de Verossimilhança e sua correção de Rao-Scott são dados respectivamente por

```
# Estatística da RV
x2rv <- 2*n*sum(phat*log(phat/p0))
x2rv
```

```
## [1] 10.56154
```

$$X_{RV}^2 = 2 \times 1.000 \times \log\left(\frac{0,84 \times 0,16}{0,80 \times 0,20}\right) = 10,56 \ ,$$

com $p$valor $0,0012$, e

```
# Estatística da RV corrigida
x2rvd <- x2rv/d
x2rvd
```

```
## [1] 0.528077
```

```
# valor-p do teste
pchisq(x2rvd, 1, lower.tail = FALSE)
```

```
## [1] 0.4674165
```

$$X_{RV}^2(d) = X_{LR}^2/d = 10,56/20 = 0,528 \ \ ,$$

com $p$valor de $0,4675$.

Como se pode notar, as estatísticas baseadas na Razão de Verossimilhança oferecem resultados semelhantes às versões correspondentes baseadas na estatística de Pearson. Em ambos os casos, as decisões baseadas nas estatísticas sem correção seriam incorretas no sentido de rejeitar a hipótese nula. Também em ambos os casos a correção de Rao-Scott produziu efeito semelhante.

O efeito de plano amostral $d = 20$ observado neste exemplo é muito grande e pouco comum na prática. Isto ocorreu neste caso porque o coeficiente de correlação intraclasse assume o valor máximo $\rho = 1$ (todos os valores dentro de um conglomerado são iguais, e portanto a homogeneidade é máxima). Na prática, as correlações intraclasse observadas são usualmente positivas mas menores que um, e portanto as estimativas de efeito de plano amostral $\widehat{d}$ correspondentes são maiores que um. Para conglomerados de tamanho médio igual a 20 ($m = 20$) como neste exemplo, os valores típicos de $\widehat{d}$ são menores que 3, tendo em correspondência correlações intraclasse estimadas positivas $\widehat{\rho} < 0,1$.

Os resultados do exemplo discutido nesta seção ilustram bem a importância de considerar o plano amostral na construção de estatísticas de teste para proporções simples, embora num caso um tanto extremo. Ilustram também um dos enfoques existentes para tratar do problema, a saber a correção de estatísticas de teste usuais (de Pearson e da Razão de Verossimilhança).

7.2.2 Estatística de Wald

Como alternativa à estatística de teste de Pearson, podemos usar a estatística de bondade de ajuste X_N^2 de Neyman. No caso de duas celas, ela se reduz a

$$X_N^2 = n\sum_{j=1}^{2}\left(\hat{p}_j - p_{0j}\right)^2/\hat{p}_j = \frac{\left(\widehat{p}-p_0\right)^2}{\hat{p}\left(1-\hat{p}\right)/n} \quad . \tag{7.6}$$

Note que a expressão de X_N^2 em (7.6) pode ser obtida substituindo-se no denominador de X_P^2 em (7.2) a proporção hipotética p_0 pela proporção estimada $\hat{p}$.

A estatística de Neyman é um caso particular da estatística de bondade de ajuste de Wald. Esta última estatística difere das estatísticas de Pearson , da Razão de Verossimilhança e de Neyman por incorporar automaticamente o plano amostral. Para o caso de duas celas, ela se reduz a

$$X_W^2 = \left(\widehat{p}-p_0\right)^2/\hat{V}_p\left(\widehat{p}\right), \tag{7.7}$$

onde $\hat{V}_p\left(\widehat{p}\right)$ é uma estimativa da variância de aleatorização de $\hat{p}$, correspondente ao plano amostral efetivamente utilizado.

O efeito do termo $\hat{V}_p\left(\widehat{p}\right)$, que aparece no denominador de X_W^2, é incorporar na estatística de bondade de ajuste o efeito do plano amostral utilizado. No caso particular de amostragem aleatória simples, usamos no lugar de $\widehat{V}_p\left(\widehat{p}\right)$ a variância $\widehat{V}_{bin}\left(\widehat{p}\right) = \widehat{p}\left(1-\widehat{p}\right)/n$. Neste caso, estatística resultante X_{bin}^2 coincide com a estatística X_N^2 de Neyman.

Para o plano amostral de conglomerados considerado no Exemplo @ref{ex:exebin}, a estatística X_W^2, sem qualquer ajuste auxiliar, já é distribuída assintoticamente como qui-quadrado com um grau de liberdade. O valor da estatística de Wald para esse exemplo é

```
x2n <- 1000* sum((phat-p0)^2/phat)
vhatp <- (phat[1]*phat[2])/m # variância usa número efetivo
# Estatística de Wald
x2w <- ((phat[1]-p0[1])^2)/vhatp
x2w
```

```
## [1] 0.5952381
```

```
pchisq(x2w,1, lower.tail = FALSE)
```

```
## [1] 0.4404007
```

$$X_W^2 = \left(0,84-0,80\right)^2/0,002743 = 0,583 \quad .$$

Observe que o valor desta estatística é bem próximo dos valores das estatísticas de Pearson e da Razão de Verossimilhança com a correção de Rao-Scott.

A estatística de Wald, pelo uso de uma estimativa apropriada da variância, reflete a complexidade do plano amostral e fornece uma estatística de teste assintoticamente válida, não necessitando que seja feito qualquer ajuste auxiliar. Esta pode ser considerada uma vantagem em relação às estatísticas com correção de Rao-Scott. Entretanto, no caso de mais de duas celas, pode haver desvantagens no uso da estatística de Wald baseada no plano amostral, devido à instabilidade nas estimativas de variância em pequenas amostras.

Reproduzimos na Tabela 7.2 os resultados para todas as estatísticas de teste consideradas até agora, para facilidade de comparação.

Nesta seção foram apresentadas as duas principais abordagens para incorporar o efeito do plano amostral na estatística de teste:

Table 7.2: Valores observados e valores-p de estatísticas de testes para os dados do Exemplo 4.4

Estatística	gl	vobs	valorp
Pearson	1	10.00	0.0016
Pearson ajustada	1	0.50	0.4795
RV	1	10.56	0.0012
RV ajustada	1	0.53	0.4674
Wald	1	0.60	0.4404

1. a metodologia de ajuste de Rao-Scott para as estatísticas de teste de Pearson e da Razão de Verossimilhança;
2. e a estatística de Wald baseada no plano amostral.

Ambas as abordagens são facilmente generalizáveis para tabelas de uma ou duas entradas com número de linhas e colunas maior que dois. Vamos considerar na próxima seção o caso geral de testes de bondade de ajuste e apresentar mais detalhes sobre as estatísticas de teste alternativas. Depois, introduziremos os testes de independência e de homogeneidade para tabelas de duas entradas. A ênfase será dada nos procedimentos baseados na estatísticas de teste de Wald baseadas no plano amostral e nas estatísticas de Pearson e da RV com os vários ajustes de Rao-Scott.

7.3 Teste para Várias Proporções

Neste seção vamos considerar extensões do problema de testes de bondade de ajuste, aumentando o número de proporções envolvidas. O caso de tabelas de duas entradas será considerado no capítulo seguinte.

A hipótese de bondade de ajuste para $J \geq 2$ celas pode ser escrita como $H_0 : p_j = p_{0j}$, $j = 1, \ldots, J$, onde $p_j = N_j/N$ são as proporções populacionais desconhecidas nas celas e p_{0j} são as proporções hipotéticas das celas. Essa hipótese pode também ser escrita, usando notação vetorial, como $H_0 : \mathbf{p} = \mathbf{p}_0$, onde $\mathbf{p} = (p_1, \ldots, p_{J-1})^{'}$ é o vetor de proporções populacionais desconhecidas e $\mathbf{p}_0 = (p_{01}, \ldots, p_{0\,J-1})^{'}$ é o vetor de proporções hipotéticas.

O vetor de estimativas consistentes das proporções das celas, baseado em n observações, é denotado por $\widehat{\mathbf{p}} = (\widehat{p}_1, \ldots, \widehat{p}_{J-1})^{'}$, onde $\widehat{p}_j = \widehat{n}_j/n$. Os $\widehat{n}_j$ são as frequências ponderadas nas celas, considerando as diferentes probabilidades de inclusão dos elementos e ajustes por não-resposta, onde os pesos amostrais são normalizados de modo que $\sum_{j=1}^{J} \widehat{n}_j = n$. Se n não for fixado de antemão, os $\widehat{p}$ serão estimadores de razões, o que é comum quando trabalhamos com subgrupos da população. Observe que apenas $J - 1$ componentes são incluídos em cada um dos vetores $\mathbf{p}$, $\mathbf{p}_0$ e $\widehat{\mathbf{p}}$, pois a soma das proporções nas J categorias é igual a 1, e portanto a proporção na J-ésima categoria é obtida por diferença.

7.3.1 Estatística de Wald Baseada no Plano Amostral

A estatística de Wald baseada no plano amostral X^2_W, para o teste da hipótese simples de bondade de ajuste, foi anteriormente introduzida no caso de duas celas como uma alternativa à estatística de Pearson ajustada. No caso de mais de duas celas, a estatística de bondade de ajuste de Wald é dada por

$$X^2_W = (\widehat{\mathbf{p}} - \mathbf{p}_0)^{'} \, \widehat{\mathbf{V}}_p^{-1} \, (\widehat{\mathbf{p}} - \mathbf{p}_0) \ , \tag{7.8}$$

onde $\widehat{\mathbf{V}}_p$ denota um estimador consistente da matriz de covariância de aleatorização verdadeira $\mathbf{V}_p$ do estimador $\widehat{\mathbf{p}}$ do vetor de proporções $\mathbf{p}$. Uma estimativa $\widehat{\mathbf{V}}_p$ pode ser obtida pelo método de linearização, usando-se por exemplo o pacote **SUDAAN**.

Sob a hipótese nula H_0, a estatística X^2_W tem distribuição assintótica qui-quadrado com $J-1$ graus de liberdade, fornecendo assim um procedimento de teste válido no caso de amostras complexas. Na prática, espera-se que X^2_W funcione adequadamente se o número de unidades primárias de amostragem selecionadas for grande e o número de celas componentes do vetor $\mathbf{p}$ for relativamente pequeno. Neste caso, podemos obter um estimador estável de $\mathbf{V}_p$. Observe que (7.7) é um caso particular de (7.8).

7.3.2 Situações Instáveis

Se o número m de unidades primárias de amostragem disponíveis for pequeno, pode ocorrer um problema de instabilidade na estimativa $\widehat{\mathbf{V}}_p$, devido ao pequeno número de graus de liberdade $f = m - H$ disponível para a estimação da variância. A instabilidade da estimativa $\widehat{\mathbf{V}}_p$ pode tornar a estatística de Wald muito liberal.

é comum contornar esta instabilidade corrigindo a estatística de Wald, mediante emprego da chamada `estatística de Wald F-corrigida` . Há duas propostas alternativas de estatísticas **F**-corrigidas de Wald. A primeira é dada por

$$F_{1.p} = \frac{f - J + 2}{f\,(J-1)} X^2_W \quad , \tag{7.9}$$

que tem distribuição assintótica de referência F com $J-1$ e $f-J+2$ graus de liberdade. A segunda é dada por

$$F_{2.p} = \frac{X^2_W}{(J-1)}, \tag{7.10}$$

que tem distribuição assintótica de referência F com $J-1$ e f graus de liberdade. No caso $J=2$, as duas correções reproduzem a estatística original.

O efeito de uma correção **F** à estatística X^2_W pode ser visualizado facilmente no caso de duas celas. Se f for pequeno, então o $p$valor de X^2_W, obtido a partir de uma distribuição **F** com 1 e f graus, é maior que o p valor obtido numa distribuição qui-quadrado com um grau de liberdade. Quando f aumenta a diferença diminui, tornando a correção desprezível, quando f for grande.

(Thomas and Rao, 1987) analisaram o desempenho das diferentes estatísticas de teste de bondade de ajuste, no caso de instabilidade. Eles verificaram que a estatística de Wald F-corrigida $F_{1.p}$ não apresentou, em geral, o melhor desempenho nesta comparação, contudo, comportou-se relativamente bem nos casos padrões, onde a instabilidade não era muito grave. As estatísticas **F**-corrigidas de Wald são bastante utilizadas na prática, e estão implementadas em pacotes para análise de dados de pesquisas amostrais complexas.

7.3.3 Estatística de Pearson com Ajuste de Rao-Scott

O exemplo introdutório serviu para mostrar que, 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. Inicialmente, vamos considerar a estatística de teste X^2_P de Pearson. Essa estatística pode ser escrita em forma matricial como

$$X^2_P = n \sum_{j=1}^{J} \left(\widehat{p}_j - p_{0j}\right)^2 / p_{0j} = n \left(\widehat{\mathbf{p}} - \mathbf{p}_0\right)' \mathbf{P}_0^{-1} \left(\widehat{\mathbf{p}} - \mathbf{p}_0\right) \tag{7.11}$$

onde $\mathbf{P}_0 = diag\left(\mathbf{p}_0\right) - \mathbf{p}_0\mathbf{p}_0^{'}$ e $\mathbf{P}_0/n$ é a matriz $(J-1)\times(J-1)$ de covariância multinomial de $\widehat{\mathbf{p}}$ sob a hipótese nula, e $diag\left(\mathbf{p}_0\right)$ representa uma matriz diagonal com elementos p_{0j} na diagonal.

A matriz de covariância $\mathbf{P}_0/n$ é uma generalização do caso $J = 2$ celas para o caso de mais de duas celas ($J > 2$). Observe que a expressão de X_P^2 tem a mesma forma da estatística de Wald, com $\mathbf{P}_0/n$ no lugar de $\widehat{\mathbf{V}}_p$. No caso de apenas duas celas, X_P^2 reduz-se à fórmula simples antes considerada $X_P^2 = \left(\widehat{p}_1 - p_{01}\right)^2 / \left[p_{01}\left(1 - p_{01}\right)/n\right]$, onde o denominador corresponde à variância da binomial sob a hipótese nula.

Para examinar a distribuição assintótica da estatística X_P^2 de Pearson, vamos generalizar os resultados anteriores, do caso de duas celas para o caso $J > 2$. Neste caso, X_P^2 é assintoticamente distribuído como uma soma ponderada $\delta_1 W_1 + \delta_2 W_2 + \ldots + \delta_{J-1} W_{J-1}$ de $J - 1$ variáveis aleatórias independentes W_j , cada uma tendo distribuição qui-quadrado com um grau de liberdade. Os pesos δ_j são os autovalores da matriz de efeito multivariado de plano amostral $\boldsymbol{\Delta} = \mathbf{P}_0^{-1}\mathbf{V}_p$, onde $\mathbf{V}_p/n$ é a matriz de covariância do estimador $\widehat{\mathbf{p}}$ do vetor de proporção $\mathbf{p}$ baseada no plano amostral verdadeiro. Tais autovalores são também chamados efeitos generalizados de plano amostral. Observe que, em geral, eles não coincidem com os efeitos univariados de plano amostral d_j.

No caso de amostragem aleatória simples, os efeitos generalizados de plano amostral δ_j são todos iguais a um, pois neste caso $\boldsymbol{\Delta} = \mathbf{I}$, matriz identidade. Neste caso, a soma $\sum_{j=1}^{J-1} \delta_j W_j$ se reduz a $\sum_{j=1}^{J-1} W_j$, cuja distribuição é χ^2 com $J - 1$ graus de liberdade. Assim, sob amostragem aleatória simples, a estatística X_P^2 é distribuída assintoticamente como qui-quadrado com $J - 1$ graus de liberdade.

No caso de plano amostral mais complexo, envolvendo estratificação e/ou conglomeração, os efeitos generalizados de plano amostral não são iguais a um. Devido aos efeitos de conglomeração, os δ_j tendem a ser maiores que um, e assim a distribuição assintótica da variável aleatória $\sum_{j=1}^{J-1} \delta_j W_j$ será diferente de uma qui-quadrado com $J - 1$ graus de liberdade. Desta forma, a estatística X_P^2 requer correções semelhantes às introduzidas no caso de duas celas. No caso geral, há mais de uma possibilidade de correção e consideraremos as `correções de primeira ordem e de segunda ordem de Rao-Scott` , desenvolvidas por (Rao and Scott, 1981). A correção de primeira ordem tem por objetivo corrigir a esperança assintótica da estatística X_P^2 de Pearson, e a de segunda ordem também envolve correção da variância. Tecnicamente, os dois ajustes são baseados nos autovalores da matriz de efeito multivariado de plano amostral estimada $\widehat{\boldsymbol{\Delta}}$.

Inicialmente, consideramos um `ajuste simples de EPA médio` à estatística X_P^2, devido a (Fellegi, 1980) e (Holt et al., 1980), e o ajuste de primeira ordem de Rao-Scott. Estes ajustes são úteis nos casos em que não é possível obter uma estimativa adequada $\widehat{\mathbf{V}}_p$ para a matriz de covariância de aleatorização. Quando esta estimativa está disponível, deve-se usar o ajuste mais preciso de segunda ordem.

O ajuste de EPA médio é baseado nos efeitos univariados de plano amostral estimados $\hat{d}_j$ das estimativas $\hat{p}_j$. O ajuste da estatística (7.11) é feito dividindo o valor observado da estatística X_P^2 de Pearson pela média $\hat{d}_{\cdot}$ dos efeitos univariados de plano amostral:

$$X_P^2\left(\hat{d}_{\cdot}\right) = X_P^2/\hat{d}_{\cdot} \tag{7.12}$$

onde $\hat{d}_{\cdot} = \sum_{j=1}^{J} \hat{d}_j/J$ é um estimador da média $\bar{d}$ dos efeitos de plano amostral desconhecidos.

Estimamos os efeitos do plano amostral por $\hat{d}_j = \hat{V}_p\left(\hat{p}_j\right) / \left(\hat{p}_j\left(1 - \hat{p}_j\right)/n\right)$, onde $\hat{V}_p\left(\hat{p}_j\right)$ é a estimativa da variância de aleatorização do estimador de proporção $\hat{p}_j$. Este ajustamento requer que estejam disponíveis as estimativas dos efeitos de plano amostral dos estimadores das proporções das J celas. A correlação intraclasse positiva fornece uma média $\hat{d}_{\cdot}$ maior que 1 e, portanto, o ajuste do EPA médio tende a remover a liberalidade de X_P^2.

O ajuste do EPA médio não corrige exatamente a esperança assintótica de X_P^2 , pois a média dos efeitos univariados de plano amostral não é igual à média dos efeitos generalizados de plano amostral. Sob a hipótese nula, a esperança assintótica de X_P^2 é $E\left(X_P^2\right) = \sum_{j=1}^{J-1} \delta_j$, logo $E\left(X_P^2/\bar{\delta}\right) = E\left(\chi^2\left(J-1\right)\right) = J - 1$, onde a média dos autovalores é $\bar{\delta} = \sum_{j=1}^{J-1} \delta_j/\left(J-1\right)$. Este raciocínio conduz ao ajuste de primeira ordem de Rao-Scott para X_P^2, dado por

$$X_P^2\left(\hat{\delta}_{\cdot}\right) = X_P^2/\hat{\delta}_{\cdot} \ , \tag{7.13}$$

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

Podemos estimar a média dos efeitos generalizados usando os efeitos univariados de plano amostral estimados, pela equação

$$(J-1)\,\hat{\delta}_{\cdot} = \sum_{j=1}^{J} \frac{\hat{p}_j}{p_{0j}}\,(1-\hat{p}_{0j})\,\hat{d}_j,$$

sem estimar os próprios autovalores. Alternativamente, $\hat{\delta}_{\cdot}$ pode ser obtido a partir da estimativa da matriz de efeitos multivariados $\hat{\mathbf{\Delta}} = n\mathbf{P}_0^{-1}\hat{\mathbf{V}}_p$, pela equação $\hat{\delta}_{\cdot} = tr\left(\hat{\mathbf{\Delta}}\right)/(J-1)$, isto é, dividindo o traço de $\hat{\mathbf{\Delta}}$ pelo número de graus de liberdade.

A estatística ajustada $X_P^2\left(\hat{\delta}_{\cdot}\right)$ só tem distribuição assintoticamente qui-quadrado com $(J-1)$ graus de liberdade se os autovalores forem iguais. Na prática, esta estatística funciona bem se a variação dos autovalores estimados for pequena. No cálculo de $X_P^2\left(\hat{\delta}_{\cdot}\right)$ só são necessários os efeitos multivariados de plano amostral dos $\hat{p}_j$ que aparecem na diagonal da matriz $\hat{\mathbf{\Delta}}$. Assim, esta estatística é adequada em análises secundárias de tabelas de contingência, se forem divulgadas as estimativas de efeito de plano amostral correspondentes. O ajuste de primeira ordem de Rao-Scott $X_P^2\left(\hat{\delta}_{\cdot}\right)$ é mais exato do que o ajuste do EPA médio da estatística $X_P^2\left(\hat{d}_{\cdot}\right)$, que é considerada uma alternativa conservadora de $X_P^2\left(\hat{\delta}_{\cdot}\right)$.

A correção de primeira ordem de Rao-Scott (7.13) é introduzida na estatística de Pearson com o objetivo de tornar a média assintótica da estatística ajustada igual ao número de graus de liberdade da distribuição de referência. Se a variação dos autovalores estimados $\hat{\delta}_j$ for grande, então será também necessária uma correção da variância de X_P^2. Isto é obtido através de uma `correção de segunda ordem de Rao-Scott`, baseada no método de (Satterthwaite, 1946). A estatística de Pearson com ajuste de Rao-Scott de segunda ordem é dada por

$$X_P^2\left(\hat{\delta}_{\cdot}, \hat{a}^2\right) = X_P^2\left(\hat{\delta}_{\cdot}\right)/\left(1+\hat{a}^2\right), \tag{7.14}$$

onde $\hat{a}^2$ é um estimador do quadrado do coeficiente de variação a^2 dos autovalores desconhecidos dado por

$$\hat{a}^2 = \sum_{j=1}^{J-1} \hat{\delta}_j^2 / \left((J-1)\,\hat{\delta}_{\cdot}^2\right) - 1 \ .$$

Um estimador da soma dos quadrados dos autovalores é dado por

$$\sum_{j=1}^{J-1} \hat{\delta}_j^2 = tr\left(\hat{\mathbf{\Delta}}^2\right) = n^2 \sum_{j=1}^{J}\sum_{k=1}^{J} \hat{V}_p^2\left(\hat{p}_j, \hat{p}_k\right)/p_{0j}p_{0k},$$

onde $\hat{V}_p\left(\hat{p}_j, \hat{p}_k\right)$ são os estimadores das covariâncias de aleatorização de $\hat{p}_j$ e $\hat{p}_k$. Os graus de liberdade também devem ser corrigidos. A estatística $X_P^2\left(\hat{\delta}_{\cdot}, \hat{a}^2\right)$ é assintoticamente qui-quadrado com graus de liberdade com ajuste de Satterthwaite dados por $gl_S = (J-1)/\left(1+\hat{a}^2\right)$.

Observe que, para o ajuste de segunda ordem, é necessária estimativa completa da matriz de variância $\hat{\mathbf{V}}_p$, enquanto que para o ajuste de primeira ordem só precisamos conhecer estimativas das variâncias $\hat{V}_p$.

Table 7.3: Vetores de proporções por classes de idade da PPV 96/97 e Contagem 96 e EPAs calculados para a PPV - Região Sudeste

idade	prop_contagem	frequência	prop_est_ppv	epa
0-14	0.2842	2516	0.2845	2.2862
15-29	0.2747	2360	0.2678	2.1867
30-44	0.2263	2018	0.2225	2.2542
45-59	0.1261	1177	0.1316	1.9906
60+	0.0860	832	0.0935	3.1632

Em situações instáveis, pode ser necessário fazer uma correção F ao ajuste de primeira ordem de Rao-Scott (7.13). A estatística F-corrigida é definida por

$$FX_P^2\left(\hat{\delta}_{\cdot}\right) = X_W^2 / \left((J-1)\,\hat{\delta}_{\cdot}\right) \quad . \tag{7.15}$$

A estatística $FX_P^2\left(\hat{\delta}_{\cdot}\right)$ tem distribuição de referência F com $J-1$ e f graus de liberdade. (Thomas and Rao, 1987) observaram que esta estatística, em situações instáveis, é melhor que a estatística sem correção de primeira ordem.

Example 7.1. Teste de bondade de ajuste para a distribuição etária da PPV 96-97 na Região Sudeste.

Vamos considerar um teste da bondade de ajuste da distribuição das idades para a Pesquisa sobre Padrões de Vida (PPV) 96/97, para os subgrupos de 0 a 14; de 15 a 29; de 30 a 44; de 45 a 59 e de 60 e mais anos de idade. As proporções correspondentes para a população foram obtidas da Contagem Populacional de 96. Na Região Sudeste, o número de estratos é $H = 15$ e o número total de conglomerados (setores) na amostra da PPV é $m = 276$ e portanto $f = m - H = 261$. As informações utilizadas neste exemplo são apresentadas na Tabela 7.3.

```
library(survey)
ppv1 <- readRDS("~\\GitHub\\adac\\data\\ppv.rds")
# cria idade categorizada
ppv1<-transform(ppv1,idatab=cut(v02a08,
c(0,14,29,44,59,200), include.lowest=T), one=1)

# Objeto de desenho da PPV
ppv.des<-svydesign(id=~nsetor,strat=~estratof,weights=~pesof,
data=ppv1,nest=TRUE)
# Considera região sudeste
ppv.se.des<-subset(ppv.des,regiao==2)
# estima proporções nas classes de IDATAB
ppv.id<-svymean(~idatab,ppv.se.des,deff=T)
vhat<-vcov(ppv.id)
ppv.id <- data.frame(ppv.id)
freq <- svyby(~one, ~idatab, ppv.se.des, unwtd.count)$counts
# matriz de variância-covariância estimada
idad_tab <- c("0-14", "15-29", "30-44","45-59","60+")
tab73 <- data.frame(idade= idad_tab, prop_contagem= c(0.2842, 0.2747, 0.2263, 0.1261, 0.0860), frequênc
knitr::kable(tab73, booktabs= TRUE, digits=c(0,4,0,4,4), align = "lcccc",
 caption = "Vetores de proporções por classes de idade da PPV 96/97 e Contagem
 96 e EPAs calculados para a PPV - Região Sudeste")
```

Os valores dos EPAs observados na PPV (coluna 5 da Tabela 7.3 mostram que o plano amostral não pode ser ignorado na análise. Queremos testar a hipótese $H_0 : \mathbf{p} = \mathbf{p}_0$ usando as estimativas de proporções

obtidas pela amostra da PPV. O vetor de proporções populacionais $\mathbf{p}_0$ foi obtido dos resultados da Contagem Populacional de 96, que é uma pesquisa censitária. Neste exemplo, vamos calcular a estatística de Pearson e suas correções, e também a estatística de Wald baseada no plano amostral. Calculamos a matriz $\widehat{\mathbf{V}}_p$ pela aplicação do método de linearização de Taylor descrito na Seção 3.3 através da fórmula (**??**) obtendo

	0-14	15-29	30-44	45-59	60+
0-14	52.274	-3.899	-5.672	-19.292	-23.411
15-29	-3.899	48.164	-29.346	-3.399	-11.520
30-44	-5.672	-29.346	43.799	-8.226	-0.556
45-59	-19.292	-3.399	-8.226	25.551	5.366
60+	-23.411	-11.520	-0.556	5.366	30.120

Para obter a estatística de Pearson (7.11), vamos calcular a matriz de covariância populacional e uma estimativa dessa matriz de covariância sob suposição de distribuição multinomial, dada por $\mathbf{P}_0/n = \frac{diag(\mathbf{p}_0) - \mathbf{p}_0\mathbf{p}_0^{'}}{8.903}$, resultando em

22.850	-8.855	-7.224	-4.025	-2.745
-8.855	22.515	-7.051	-3.929	-2.680
-7.224	-7.051	19.666	-3.205	-2.186
-4.025	-3.929	-3.205	12.378	-1.218
-2.745	-2.680	-2.186	-1.218	8.829

Para obter os diversos ajustes desta estatística precisamos usar os valores dos EPAs, listados na coluna 5 da Tabela 7.3. Estes valores foram obtidos através do pacote SUDAAN. Para obter as diferentes correções da estatística de Pearson, precisamos calcular as seguintes quantidades:

```
dhat <- tab73$epa
dhatdot <- mean(dhat)
dhatdot
```

```
## [1] 2.376168
```

$$\hat{d}_{\cdot} = \sum_{j=1}^{5} \hat{d}_j/5 = 2,376 \ ,$$

```
phat <- tab73$prop_est_ppv
deltahatdot <- sum(phat*(1-p0)*dhat/(4*p0))
deltahatdot
```

```
## [1] 2.459607
```

$$\hat{\delta}_{\cdot} = \sum_{j=1}^{5} \frac{\hat{p}_j}{4p_{0j}} \left(1 - \hat{p}_{0j}\right) \hat{d}_j = 2,457 \ ,$$

```
nppv <- sum (freq)
raz <- matrix(NA,5,5)
for(i in 1:5){
  for(j in 1:5){
    raz[i,j]<- vhat[i,j]^2/(p0[i]*p0[j])
  }
}
ahat2_1 <- nppv^2* sum(raz)/(4*deltahatdot^2)
ahat2_1
```

```
## [1] 1.249524
```

$$1+\hat{a}^2 = 8903^2 \sum_{j=1}^{5}\sum_{k=1}^{5}\left(\hat{V}_p^2\left(\hat{p}_j,\hat{p}_k\right)/p_{0j}p_{0k}\right) / \left(4\times 2,457^2\right) = 1,253 \ .$$

Podemos então calcular a estatística X_P^2 de Pearson usando (7.11), resultando em

```
x2p <-nppv* sum((phat -p0)^2/p0)
x2p
```

```
## [1] 11.50719
```

$$X_P^2 = 11,64$$

com 4 g.l. e um $p$valor $0,020$.

A estatística de Pearson com ajustamento de EPA médio é calculada usando (7.12), resultando em

```
x2p/dhatdot
```

```
## [1] 4.842751
```

$$X_P^2\left(\hat{d}_{.}\right) = 11,64/2,376 = 4,901$$

com 4 g.l. e um $p$valor $0,298$.

A estatística de Pearson com ajustamento de Rao-Scott de primeira ordem, dada por (7.13), resulta em

```
x2p/deltahatdot
```

```
## [1] 4.678467
```

$$X_P^2\left(\hat{\delta}_{.}\right) = 11,64/2,457 = 4,74$$

com 4 g.l. e um $p$valor $0,315$.

O ajustamento de Rao-Scott de primeira ordem F-corrigido para a estatística de Pearson, dado por (7.15), resulta em

```
J <- length(phat)
x2p/ ((J-1)*deltahatdot)
```

```
## [1] 1.169617
```

$$FX_P^2\left(\hat{\delta}_{.}\right) = 4,74/4 = 1,85$$

com 4 e 261 g.l e um $p$valor $0,318$.

O ajustamento de Rao-Scott de segunda ordem para a estatística de Pearson, dado por (7.14), resulta em

```
(x2p/deltahatdot)/ahat2_1
```

```
## [1] 3.744199
```

$$X_P^2\left(\hat{\delta}_{.},\hat{a}^2\right) = 4,74/1,253 = 3,784$$

com $4/1,253 = 3,19$ g.l. e $p$valor $0,314$.

A estatística de Wald baseada no plano amostral (veja equação (7.8) resulta em

```
phatv <- matrix(phat[-5], ncol=1)
p0v <- matrix(p0[-5], ncol=1)
vhat0 <- vhat[-5,-5]
x2w <-  t(phatv-p0v)%*% solve(vhat0)%*%(phatv-p0v)
x2w
```

```
##          [,1]
## [1,] 5.742022
```

$$X^2_W = 5,691$$

com 4 g.l. e um $p$valor $0,223$.

As estatísticas F-corrigidas de Wald, definidas em (7.9) e (7.10), resultam em

```
# número de estratos da PPV-sudeste
nestrat <- length(unique(ppv1$estratof[ppv1$regiao == 2]))
# número de setores da PPV-sudeste
nsetor <- length(unique(ppv1$nsetor[ppv1$regiao == 2]))
f <- nsetor -nestrat
f1p <- ((f-J+2)/(f*(J-1)))*x2w
f1p
```

```
##          [,1]
## [1,] 1.419005
```

$$F_{1.p} = \frac{261 - 5 + 2}{261 \times 4} \times 5,690661 = 1,406$$

com 4 e 259 g.l. e um $p$valor $0,232$, e

```
f2p <- x2w/(J-1)
f2p
```

```
##          [,1]
## [1,] 1.435505
```

$$F_{2.p} = 5,691/4 \ = 1,423$$

com 4 e 261 gl e um $p$valor $0,228$.

A Tabela 7.4 resume os valores das diversas estatísticas de teste calculadas, bem como das informações comparativas com as respectivas distribuições de referência.

Table 7.4: Valores e valores-p de estatísticas alternativas de teste

Estatística	Tipo	Valor	Distribuição	valor-p
X^2_P	Adequada para IID	11.640	$\chi^2(4)$	0.020
$X^2_P\left(\hat{d}.\right)$	Ajustes e	4.901	$\chi^2(4)$	0.298
$X^2_P\left(\hat{\delta}.\right)$	correções da	4.740	$\chi^2(4)$	0.315
$FX^2_P\left(\hat{\delta}.\right)$	Estatística	1.850	$F\,(4;261)$	0.318
$X^2_P\left(\hat{\delta}., \hat{a}^2\right)$	X^2_P	3.784	$\chi^2(3,19)$	0.314
X^2_W	Baseadas no	5.691	$\chi^2(4)$	0.223
$F_{1.p}$	plano	1.406	$F\,(4;259)$	0.232

Table 7.5: Estimativas das proporções nas classes

idade	mean	SE	deff
0 a 14 anos	0.2845	0.0072	2.286
15 a 29 anos	0.2678	0.0069	2.187
30 a 44 anos	0.2225	0.0066	2.254
45 a 59 anos	0.1316	0.0051	1.991
60 anos e mais	0.0935	0.0055	3.163

Estatística	Tipo	Valor	Distribuição	valor-p
$F_{2.p}$	amostral	1.423	$F\,(4; 261)$	0.228

Examinando os resultados da Tabela 7.4, verificamos que o teste clássico de Pearson rejeita a hipótese nula H_0 no nível $\alpha = 5\%$, diferentemente de todos os outros testes. Os valores das estatísticas com ajustes de Rao-Scott (com ou sem correção F) são semelhantes e parecem corrigir exageradamente o p-valor dos testes. A estatística de Wald baseada no plano amostral e suas correções F, que têm valores quase iguais, produzem uma correção menor no p-valor do teste. Nesse exemplo, como o número de graus de liberdade (dado pelo número de unidades primárias na amostra menos o número de estratos)$f = m - H = 261$ é grande, a correção F tem pouco efeito, tanto nas estatísticas com ajustes de primeira e segunda ordem de Rao-Scott, como na estatística Wald.

7.4 Laboratório de R

Exemplo 7.1 pode ser substituído por: Criar variável ITAB (Não aparece)

```
ppv.des<-svydesign(id=~nsetor,strat=~estratof,weights=~pesof,
data=ppv1,nest=TRUE)
ppv.se.des<-subset(ppv.des,regiao==2)
```

```
ppv.id<-svymean(~idatab,ppv.se.des,deff=T)
vhvat<-vcov(ppv.id)
```

```
library(xtable)
fr_ppv_id<- data.frame(ppv.id)
row.names(fr_ppv_id) <- NULL
fr_ppv_id <- cbind(idade= c("0 a 14 anos","15 a 29 anos", "30 a 44 anos","45 a 59 anos", "60 anos e mais
knitr::kable(fr_ppv_id, booktabs= TRUE, digits=c(0,4,4,3), caption="Estimativas das proporções nas class
```

Estatística de Wald calculada a partir da fórmula (7.8)

```
#Vetor de proporções estimadas
phat<-coefficients(ppv.id)
# Vetor de proporções obtido na Contagem Populacional de 1996
p0<-c(.2842,.2774,.2263,.1261,.086)
# Estatística de Wald
x2_w<-matrix((phat-p0)[-5],nrow=1)%*%solve(vhat[-5,-5])%*%
matrix((phat-p0)[-5],ncol=1)
x2_w
```

```
##          [,1]
## [1,] 5.742022
```

```
#Cálculo do p-valor
round(pchisq(x2_w,4,lower.tail=FALSE),digits=3)
```

```
##       [,1]
## [1,] 0.219
```

Estatística de Pearson calculada a partir da fórmula 7.11

```
n<-8903
P0<-diag(p0)-matrix(p0,ncol=1)%*%matrix(p0,nrow=1)
x2_p<-n*matrix((phat-p0)[-5],nrow=1)%*%solve(P0[-5,-5])%*%
matrix((phat-p0)[-5],ncol=1)
x2_p
```

```
##          [,1]
## [1,] 11.50719
```

Cálculo do valor-p:

```
round(pchisq(x2_p,4,lower.tail=FALSE),digits=3)
```

```
##       [,1]
## [1,] 0.021
```

Chapter 8

Testes em Tabelas de Duas Entradas

8.1 Introdução

8.2 Tabelas 2x2

8.2.1 Teste de Independência

8.2.2 Teste de Homogeneidade

8.2.3 Efeitos de Plano Amostral nas Celas

8.3 Tabelas de Duas Entradas (Caso Geral)

8.3.1 Teste de Homogeneidade

8.3.2 Teste de Independência

8.3.3 Estatística de Wald Baseada no Plano Amostral

8.3.4 Estatística de Pearson com Ajuste de Rao-Scott

8.4 Laboratório de R

8.5 [1] ``stra'' ``psu'' ``pesopes'' ``informal'' ``sx'' ``id''

8.6 [7] ``ae'' ``ht'' ``re'' ``um''

8.7 stra psu pesopes informal sx id ae

8.8 ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric''

8.9 ht re um

8.10 ``numeric'' ``numeric'' ``numeric''

Chapter 9

Estimação de densidades

9.1 Introdução

Chapter 10

Modelos Hierárquicos

10.1 Introdução

Chapter 11

Não-Resposta

11.1 Introdução

Chapter 12

Diagnóstico de ajuste de modelo

12.1 Introdução

Chapter 13

Agregação vs. Desagregação

13.1 Introdução

13.2 Modelagem da Estrutura Populacional

13.3 Modelos Hierárquicos

13.4 Análise Desagregada: Prós e Contras

Chapter 14

Pacotes para Analisar Dados Amostrais

14.1 Introdução

14.2 Pacotes Computacionais

Referências

Bibliography

Fellegi, I. P. (1980). Approximate tests of independence and goodness-of-fit based on stratified multistage samples. Journal of the American Statistical Association, 75:261--268.

Holt, D., Scott, A., and Ewings, P. D. (1980). Chi-squared tests with survey data. Journal of the Royal Statistical Society A, 143:303--320.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two way tables. Journal of the American Statistical Association, 76:221--230.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics, 2:110--114.

Thomas, D. R. and Rao, J. N. K. (1987). Small-sample comparison of level and power for simple goodness-of-fit statistics under cluster sampling. Journal of the American Statistical Association, 82:630--636.