

Análise de Dados Amostrais Complexos

Djalma Pessoa e Pedro Nascimento Silva

2017-09-05

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 regioao 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.Padiao 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,] 0.01313725	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] 0.1402485	18
4.18	[1] 0.2827575	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	24
7.2	Teste para uma Proporção	24
7.3	[1] 10	24
7.4	[1] 20	24
7.5	[1] 0.5	24
7.6	[1] 0.4795001	24
7.7	[1] 10.56154	24
7.8	[1] 0.528077	24
7.9	[1] 0.4674165	24
7.10	[1] 0.5952381	24
7.11	[1] 0.4404007	24
7.12	Teste para Várias Proporções	24
7.13	[1] 2.376168	24
7.14	[1] 2.459607	24
7.15	[1] 1.249524	24
7.16	[1] 11.50719	24
7.17	[1] 4.842751	24
7.18	[1] 4.678467	24
7.19	[1] 1.169617	24
7.20	[1] 3.744199	24
7.21	[,1]	24
7.22	[1,] 5.742022	24
7.23	[,1]	24
7.24	[1,] 1.419005	24
7.25	[,1]	24
7.26	[1,] 1.435505	24
7.27	Laboratório de R	24
7.28	[,1]	24
7.29	[1,] 5.742022	24
7.30	[,1]	24
7.31	[1,] 0.219	24
7.32	[,1]	24
7.33	[1,] 11.50719	24
7.34	[,1]	24
7.35	[1,] 0.021	24

8	Testes em Tabelas de Duas Entradas	25
8.1	Introdução	26
8.2	Tabelas 2x2	26
8.3	Tabelas de Duas Entradas (Caso Geral)	26
8.4	Laboratório de R	26
8.5	[1] ``stra'' ``psu'' ``pesopes'' ``informal'' ``sx'' ``id''	26
8.6	[7] ``ae'' ``ht'' ``re'' ``um''	26
8.7	stra psu pesopes informal sx id ae	26
8.8	``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric'' ``numeric''	26
8.9	ht re um	26
8.10	``numeric'' ``numeric'' ``numeric''	26
8.11	mean SE	26
8.12	sx1 0.65708 0.0056	26
8.13	sx2 0.34292 0.0056	26
8.14	mean SE	26
8.15	re1 0.15546 0.0069	26
8.16	re2 0.58356 0.0082	26
8.17	re3 0.26098 0.0096	26
8.18	mean SE	26
8.19	ae1 0.31304 0.0095	26
8.20	ae2 0.31972 0.0071	26
8.21	ae3 0.36725 0.0105	26
8.22	ht1 ht2 ht3	26
8.23	0.2103714 0.6148881 0.1747405	26
8.24	\$names	26
8.25	[1] ``ht1'' ``ht2'' ``ht3''	26
8.26	26
8.27	\$var	26
8.28	ht1 ht2 ht3	26
8.29	ht1 3.666206e-05 -3.322546e-05 -3.436592e-06	26
8.30	ht2 -3.322546e-05 6.758652e-05 -3.436106e-05	26
8.31	ht3 -3.436592e-06 -3.436106e-05 3.779765e-05	26
8.32	26
8.33	\$statistic	26
8.34	[1] ``mean''	26
8.35	26
8.36	\$class	26
8.37	[1] ``svystat''	26
8.38	ht1 ht2 ht3	26
8.39	ht1 3.666206e-05 -3.322546e-05 -3.436592e-06	26
8.40	ht2 -3.322546e-05 6.758652e-05 -3.436106e-05	26
8.41	ht3 -3.436592e-06 -3.436106e-05 3.779765e-05	26
8.42	ht1 ht2 ht3	26
8.43	ht1 3.666206e-05 -3.322546e-05 -3.436592e-06	26
8.44	ht2 -3.322546e-05 6.758652e-05 -3.436106e-05	26
8.45	ht3 -3.436592e-06 -3.436106e-05 3.779765e-05	26
8.46	mean SE DEff	26
8.47	re1 0.1554555 0.0068977 2.3611	26
8.48	re2 0.5835630 0.0082001 1.8027	26
8.49	re3 0.2609815 0.0096130 3.1216	26
8.50	re	26
8.51	sx 1 2 3	26
8.52	1 0.073 0.388 0.196	26
8.53	2 0.082 0.195 0.065	26

8.54	sx re1 re2 re3 se.re1 se.re2 se.re3	26
8.55	1 1 0.1110831 0.5908215 0.2980955 0.005726888 0.01025759 0.01112131	26
8.56	2 2 0.2404788 0.5696548 0.1898663 0.012502636 0.01193753 0.01114102	26
8.57	mean SE DEff	26
8.58	I((sx == 1 & re == 1) * 1) 0.0729904 0.0038343 1.4156	26
8.59	re	26
8.60	sx 1 2 3	26
8.61	1 0.07299044 0.38821684 0.19587250	26
8.62	2 0.08246505 0.19534616 0.06510900	26
8.63	re	26
8.64	sx 1 2 3	26
8.65	1 7.299044 38.821684 19.587250	26
8.66	2 8.246505 19.534616 6.510900	26
8.67	[1] 1.642161	26
8.68	26
8.69	Pearson's Chi-squared test	26
8.70	26
8.71	data: tab.amo	26
8.72	X-squared = 227.03, df = 2, p-value < 2.2e-16	26
9	Estimação de densidades	27
9.1	Introdução	27
10	Modelos Hierárquicos	29
10.1	Introdução	29
11	Não-Resposta	31
11.1	Introdução	31
12	Diagnóstico de ajuste de modelo	33
12.1	Introdução	33
13	Agregação vs. Desagregação	35
13.1	Introdução	35
13.2	Modelagem da Estrutura Populacional	35
13.3	Modelos Hierárquicos	35
13.4	Análise Desagregada: Prós e Contras	35
14	Pacotes para Analisar Dados Amostrais	37
14.1	Introdução	37
14.2	Pacotes Computacionais	37
15	Placeholder	39

Prefácio

Uma preocupação básica de toda instituição produtora de informações estatísticas é com a utilização "correta" de seus dados. Isso pode ser interpretado de várias formas, algumas delas com reflexos até na confiança do público e na própria sobrevivência do órgão. Do nosso ponto de vista, como técnicos da área de metodologia do IBGE, enfatizamos um aspecto técnico particular, mas nem por isso menos importante para os usuários dos dados.

A revolução da informática com a resultante facilidade de acesso ao computador, criou condições extremamente favoráveis à utilização de dados estatísticos, produzidos por órgãos como o IBGE. Algumas vezes esses dados são utilizados para fins puramente descritivos. Outras vezes, porém, sua utilização é feita para fins analíticos, envolvendo a construção de modelos, quando o objetivo é extrair conclusões aplicáveis também a populações distintas daquela da qual se extraiu a amostra. Neste caso, é comum empregar, sem grandes preocupações, pacotes computacionais padrões disponíveis para a seleção e ajuste de modelos. É neste ponto que entra a nossa preocupação com o uso adequado dos dados produzidos pelo IBGE.

O que torna tais dados especiais para quem pretende usá-los para fins analíticos? Esta é a questão básica que será amplamente discutida ao longo deste texto. A mensagem principal que pretendemos transmitir é que certos cuidados precisam ser tomados para utilização correta dos dados de pesquisas amostrais como as que o IBGE realiza.

O que torna especiais dados como os produzidos pelo IBGE é que estes são obtidos através de pesquisas amostrais complexas de populações finitas que envolvem: **probabilidades distintas de seleção, estratificação e conglomeração das unidades, ajustes para compensar não-resposta e outros ajustes**. Os pacotes tradicionais de análise ignoram estes aspectos, podendo produzir estimativas incorretas tanto dos parâmetros como para as variâncias destas estimativas. Quando utilizamos a amostra para estudos analíticos, as opções disponíveis nos pacotes estatísticos usuais para levar em conta os pesos distintos das observações são apropriadas somente para observações independentes e identicamente distribuídas (IID). Além disso, a variabilidade dos pesos produz impactos tanto na estimação pontual quanto na estimação das variâncias dessas estimativas, que sofre ainda influência da estratificação e conglomeração.

O objetivo deste livro é analisar o impacto das simplificações feitas ao utilizar procedimentos e pacotes usuais de análise de dados, e apresentar os ajustes necessários desses procedimentos de modo a incorporar na análise, de forma apropriada, os aspectos aqui ressaltados. Para isto serão apresentados exemplos de análises de dados obtidos em pesquisas amostrais complexas, usando pacotes clássicos e também pacotes estatísticos especializados. A comparação dos resultados das análises feitas das duas formas permitirá avaliar o impacto de ignorar o plano amostral na análise dos dados resultantes de pesquisas amostrais complexas.

Agradecimentos

A elaboração de um texto como esse não se faz sem a colaboração de muitas pessoas. Em primeiro lugar, agradecemos à Comissão Organizadora do SINAPE por ter propiciado a oportunidade ao selecionar nossa proposta de minicurso. Agradecemos também ao IBGE por ter proporcionado as condições e os meios usados

para a produção da monografia, bem como o acesso aos dados detalhados e identificados que utilizamos em vários exemplos.

No plano pessoal, agradecemos a Zélia Bianchini pela revisão do manuscrito e sugestões que o aprimoraram. Agradecemos a Marcos Paulo de Freitas e Renata Duarte pela ajuda com a computação de vários exemplos. Agradecemos a Waldecir Bianchini, Luiz Pessoa e Marinho Persiano pela colaboração na utilização do processador de textos. Aos demais colegas do Departamento de Metodologia do IBGE, agradecemos o companheirismo e solidariedade nesses meses de trabalho na preparação do manuscrito.

Finalmente, agradecemos a nossas famílias pela aceitação resignada de nossas ausências e pelo incentivo à conclusão da empreitada.

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_raz212) * c(-1, 1)`: Recycling array

Chapter 7

Testes de Qualidade de Ajuste

7.1 Introdução

7.2 Teste para uma Proporção

7.2.1 Correção de Estatísticas Clássicas

7.3 [1] 10

7.4 [1] 20

7.5 [1] 0.5

7.6 [1] 0.4795001

7.7 [1] 10.56154

7.8 [1] 0.528077

7.9 [1] 0.4674165

7.9.1 Estatística de Wald

7.10 [1] 0.5952381

7.11 [1] 0.4404007

7.12 Teste para Várias Proporções

7.12.1 Estatística de Wald Baseada no Plano Amostral

7.12.2 Situações Instáveis

7.12.3 Estatística de Pearson com Ajuste de Rao-Scott

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

Chapter 15

Placeholder

Bibliography