A Minimal Book Example

Yihui Xie

2016-12-19

Contents

Chapter 1

Prerequisites

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

For now, you have to install the development versions of **bookdown** from Github:

```
devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need to install XeLaTeX.

```
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
```

```
source("C:\\Users\\anthonyd\\Documents\\GitHub\\adac\\data\\ppv1.R")
```

Prefácio

Uma preocupação básica de toda instituição produtora de informações estatísticas é com a utilização ''correta" de seus dados. Isso pode ser intrepretado 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 2

Introdução

2.1 Motivação

Este livro trata de problema de grande importância para os usuários de dados obtidos através de pesquisas amostrais por agências produtoras de informações estatísticas. Tais dados são comumente utilizados em análises descritivas envolvendo o cálculo de estimativas para totais, proporções, médias e razões, nas quais, em geral, são devidamente considerados os pesos distintos das observações e o planejamento da amostra que lhes deu origem.

Outro uso destes dados, denominado secundário, é a construção e ajuste de modelos, feita geralmente por analistas que trabalham fora das agências produtoras dos dados. Neste caso, o foco é, essencialmente, estabelecer a natureza de relações ou associações entre variáveis. Para isto, a estatística clássica conta com um arsenal de ferramentas de análise, já incorporado aos principais pacotes estatísticos disponíveis. O uso destes pacotes se faz, entretanto, sob condições que não refletem a complexidade usualmente envolvida nas pesquisas amostrais de populações finitas. Em geral, partem de hipóteses básicas que só são válidas quando os dados são obtidos através de amostras aleatórias simples com reposição (AASC). Tais pacotes estatísticos não consideram os seguintes aspectos relevantes no caso de amostras complexas:

i.) **probabilidades distintas de seleção das unidades**;

ii.) **conglomeração das unidades**;

iii.) **estratificação**;

iv.) **calibração ou imputação para não-resposta e outros ajustes**.

As estimativas pontuais de parâmetros da população ou de modelos são influenciadas por pesos distintos das observações. Além disso, as estimativas de variância (ou da precisão dos estimadores) são influenciadas pela conglomeração, estratificação e pesos, ou no caso de não resposta, também por eventual imputação de dados faltantes. Ao ignorar estes aspectos, os pacotes tradicionais de análise podem produzir estimativas incorretas das variâncias das estimativas pontuais.

A seguir vamos apresentar um exemplo de uso de dados de uma pesquisa amostral real para ilustrar como os pontos i) a iv) acima mencionados afetam a inferência sobre quantidades descritivas populacionais tais como médias, proporções, razões e totais.

2.2 Impacto dos pesos da amostra da PPV

Os dados deste exemplo são relativos à distribuição dos pesos na amostra da Pesquisa sobre Padrões de Vida (PPV), realizada pelo IBGE nos anos 1996-97. (**?**) descrevem resumidamente a Pesquisa sobre Padrões de

Vida (PPV), que foi realizada nas Regiões Nordeste e Sudeste do País, considerando 10 estratos geográficos, a saber: Região Metropolitana de Fortaleza, Região Metropolitana de Recife, Região Metropolitana de Salvador, restante da área urbana do Nordeste, restante da área rural do Nordeste, Região Metropolitana de Belo Horizonte, Região Metropolitana do Rio de Janeiro, Região Metropolitana de São Paulo, restante da área urbana do Sudeste e restante da área rural do Sudeste.

O plano amostral empregado na seleção da amostra da PPV foi de dois estágios, com estratificação das unidades primárias de amostragem (no caso os setores censitários da base geográfica do IBGE conforme usada para o Censo Demográfico de 1991), seleção destes setores com probabilidade proporcional ao tamanho, e seleção aleatória das unidades de segundo estágio (domicílios). O tamanho da amostra para cada estrato geográfico foi fixado em 480 domicílios, e o número de setores selecionados foi fixado em 60, com 8 domicílios selecionados em cada setor. A exceção ficou por conta dos estratos que correspondem ao restante da área rural de cada Região, onde foram selecionados 30 setores e 16 domicílios por setor, em função da dificuldade de acesso a esses setores, o que implicaria em aumento de custo da coleta.

Os setores de cada um dos 10 estratos geográficos foram subdivididos em 3 estratos de acordo com a renda média mensal do chefe do domicílio por setor, perfazendo um total de 30 estratos geográficos versus renda. Em seguida foi feita uma alocação proporcional, com base no número de domicílios particulares permanentes ocupados do estrato de renda no universo de cada estrato geográfico, obtidos pelo Censo de 1991. No final foram obtidos 554 setores na amostra, distribuídos tal como revela a Tabela **??**.

Estrato Geográfico	População	Amostra
1 RM Fortaleza	2.263	62
2-RM Recife	2.309	61
3-RM Salvador	2.186	61
4-Restante Nordeste Urbano	15.057	61
5-Restante Nordeste Rural	23.711	33
6-RM Belo Horizonte	3.283	62
7-RM Rio de Janeiro	10.420	61
8-RM São Paulo	14.943	61
9-Restante Sudeste Urbano	25.855	61
10-Restante Sudeste Rural	12.001	31
Total	112.001	554

Table 2.2: Número de setores na população e na amostra, por estrato geográfico

Estrato Geográfico	Número de setores	
	População	Amostra
1-RM Fortaleza	2.263	62
2-RM Recife	2.309	61
3-RM Salvador	2.186	61
4-Restante Nordeste Urbano	15.057	61
5-Restante Nordeste Rural	23.711	33
6-RM Belo Horizonte	3.283	62
7-RM Rio de Janeiro	10.420	61
8-RM São Paulo	14.931	61
9-Restante Sudeste Urbano	25.855	61
10-Restante Sudeste Rural	12.001	31
Total	112.016	554

A Tabela **??** apresenta um resumo das distribuições dos pesos amostrais para as Regiões Nordeste (5 estratos