Utilização de dados públicos dos Censos Demográficos brasileiros

Pedro Luis do Nascimento Silva
Escola Nacional de Ciências Estatística - ENCE

Ricardo Luiz Cardoso e Sonia Oliveira

6º Seminário de Metodologia do IBGE – SMI2017

Rio de Janeiro, novembro de 2017

Nesta sessão

1. Dados da **amostra** do Censo Demográfico: o que têm de diferente.
2. **Estratificação**, **conglomeração** e **pesos amostrais** e seus efeitos na inferência e na análise.
3. Os **métodos** e **ferramentas** disponíveis para analisar dados de amostras complexas.
4. Analisando dados da amostra do Censo Demográfico 2010 usando pacotes do **sistema R**.

Alguns Conceitos Fundamentais

População alvo: conjunto de todas as unidades para as quais gostaríamos de obter informações.

População de pesquisa: conjunto de todas as unidades para as quais a pesquisa vai de fato tentar obter informações.

Unidade: um único indivíduo ou elemento a ser medido ou observado na pesquisa / censo.

Censo: coleta informações sobre características de interesse **de todas as unidades de uma população de pesquisa**, usando conceitos, métodos e procedimentos bem definidos.

Censos Demográficos no Brasil

Desde 1960, Censos Demográficos brasileiros empregam **amostragem** para coleta de parte das informações de interesse, mediante o **uso de dois questionários**:

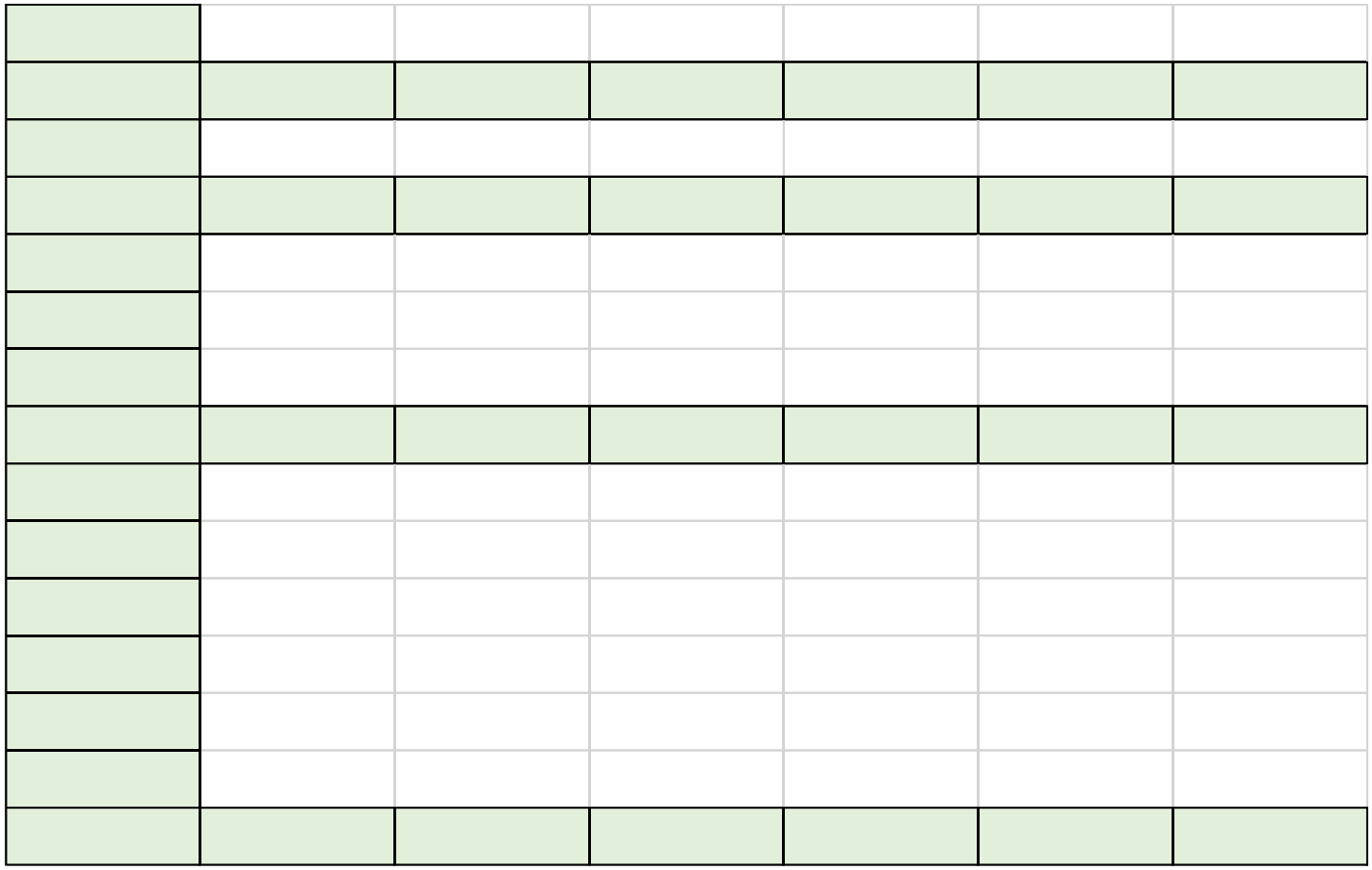
- Censo tradicional (**questionário** curto ou **básico**); e
- Pesquisa socioeconômica por amostragem (**questionário** longo, ou **da amostra**).

Todas as perguntas do questionário curto estão contidas também no questionário longo.

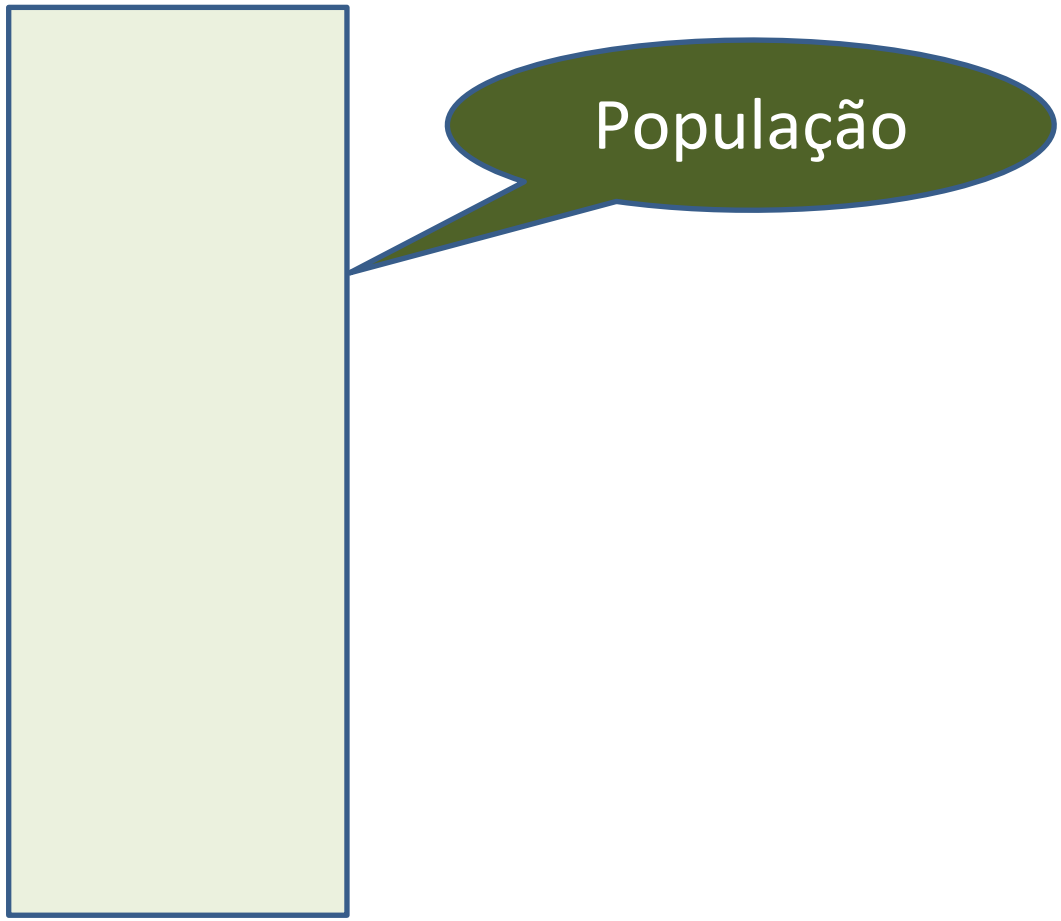
União das respostas às perguntas comuns aos dois questionários fornece o '**Conjunto Universo**'.

Questionários curto e longo no Censo Demográfico

Unidades ↓ ← Variáveis →



Censo Demográfico – Conjunto Universo



Censo Demográfico – Conjunto Universo

'**Retrato completo**' da população.

'**Alta resolução**' para **poucas variáveis**: contagens da população para pequenos domínios:

- Arquivo Agregado de Setores; ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/Agregados_por_Setores_Censitarios/
- Grade (https://censo2010.ibge.gov.br/).

Microdados (de pessoas e domicílio) **não disponíveis** para uso público por razões de **proteção de confidencialidade**.

Acessíveis apenas na **sala de acesso a dados restritos**, mediante projeto.

Notação

População de pesquisa

$U = \{1, 2, \ldots, i, \ldots, N\}$ → conjunto de N rótulos distintos

N = **tamanho da população** de pesquisa

i → rótulo para unidade genérica da população

y → variável de pesquisa / de interesse

y_i → valor da variável y para unidade i

Dados individuais num Censo

Unidade	Variável y
1	y_1
2	y_2
⋮	⋮
i	y_i
⋮	⋮
N	y_N

➔ Mostrar arquivo **agregado de setores**.

Dados individuais num Censo

	A	V	W	X	Y	Z
1	Cod_setor	V001	V002	V003	V004	V005
2	330010005000001	156	409	2,62	1,94	2.356,80
3	330010005000002	57	143	2,51	1,50	2.040,47
4	330010005000003	343	1055	3,08	2,19	2.687,80
5	330010005000004	72	219	3,04	3,25	3.026,67
6	330010005000005	212	709	3,34	1,94	1.124,82
7	330010005000006	249	740	2,97	1,91	1.283,27
8	330010005000007	361	1186	3,29	2,43	832,03
9	330010005000008	261	829	3,18	2,21	1.298,85
10	330010005000009	250	664	2,66	1,61	1.859,59
11	330010005000010	195	534	2,74	2,63	2.112,76
12	330010005000011	302	905	3,00	2,55	1.319,23
13	330010005000012	311	990	3,18	2,52	1.139,13
14	330010005000013	225	677	3,01	3,17	1.414,68
15	330010005000014	201	607	3,02	1,61	1.047,58

Parâmetros de interesse

Total populacional → $$Y = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i$$

Média populacional → $$\bar{Y} = Y / N = \sum_{i \in U} y_i / N$$

Variância populacional → $$\sigma_y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{Y})^2$$

Exemplificando parâmetros de interesse

```
# Calcula resumos para estado do RJ - Domicílios Particulares Permanentes
total_domicilios <- summarise(setorrj_dat,
                              domicilios=sum(v001, na.rm=TRUE),
                              moradores =sum(v002, na.rm=TRUE))
ftot(total_domicilios)
```

```
##   domicilios    moradores
##    5.243.011   15.923.940
```

```
# Média de Domicílios Particulares Permanentes por Setor
media_dpp_porsetor <- summarise(setorrj_dat,
                                domicilios=mean(v001, na.rm=TRUE))
```

```
##   domicilios
## 1     189,38
```

```
# Média de Habitantes por Domicílio Particular Permanente
fprop(mutate(total_domicilios, nmorpordpp=moradores/domicilios) %>%
        select(nmorpordpp))
```

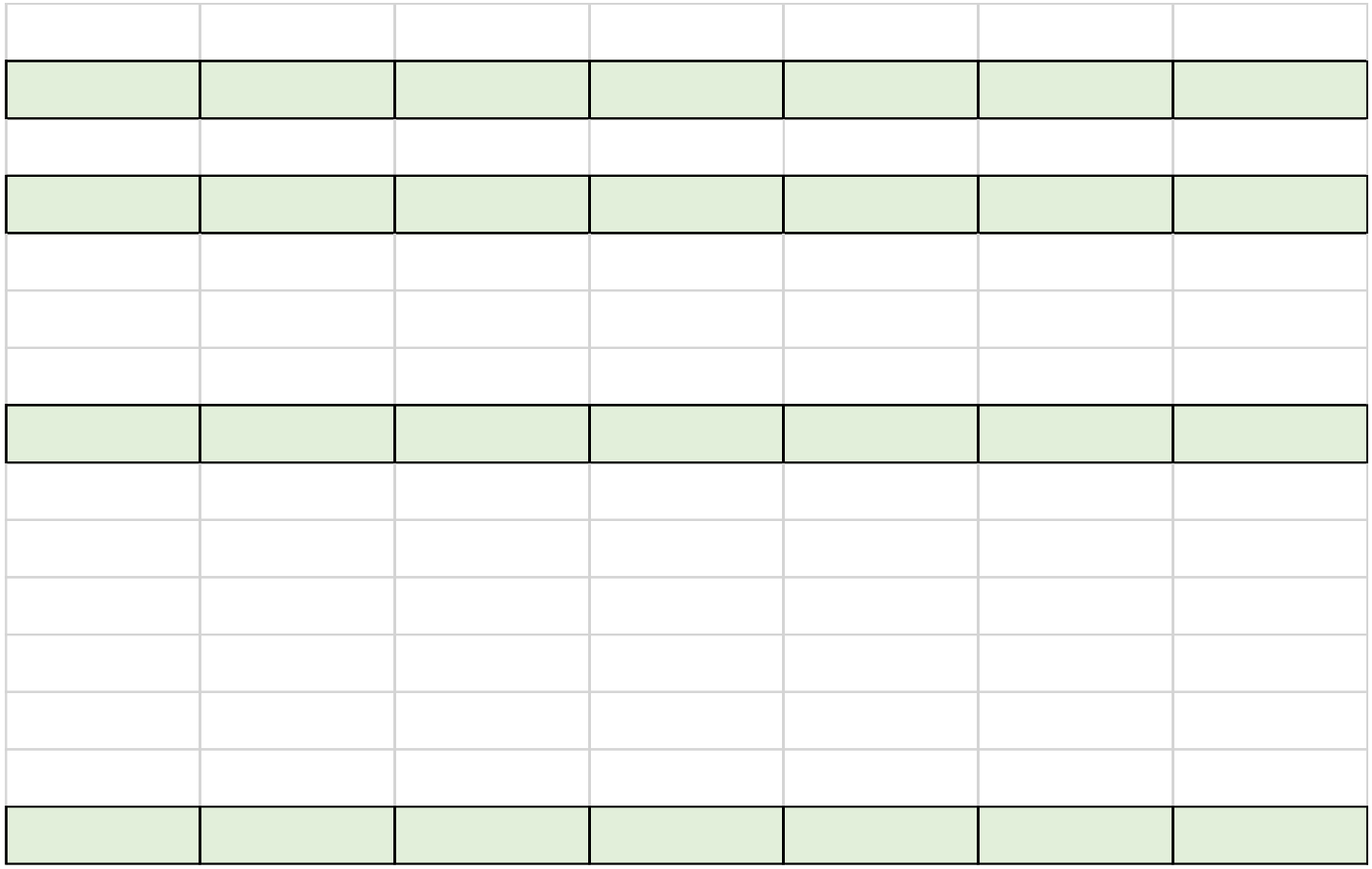
```
##   nmorpordpp
## 1       3,04
```

Por que usamos amostragem no Censo?

1. Para reduzir custo da coleta de informações.
2. Para reduzir carga de coleta sobre a população de pesquisa.
3. Para proteger a confidencialidade das informações.

Censo Demográfico – Conjunto Amostra

Unidades ↓ ← Variáveis →



Censo Demográfico – Conjunto Amostra

'**Média resolução**' para conjunto maior de variáveis:

- Estimativas de indicadores socioeconômicos para domínios geográficos médios tais como **áreas de ponderação** e **municípios**.
- Variáveis disponíveis são pesquisadas apenas no **questionário longo**.
- Dados obtidos por **amostragem probabilística**.

Microdados da amostra **disponíveis para uso público**, mas menor área geográfica identificável é '**área de ponderação**'.

Área de ponderação

Menor **unidade territorial** para divulgação de resultados da pesquisa por amostra (questionário longo).

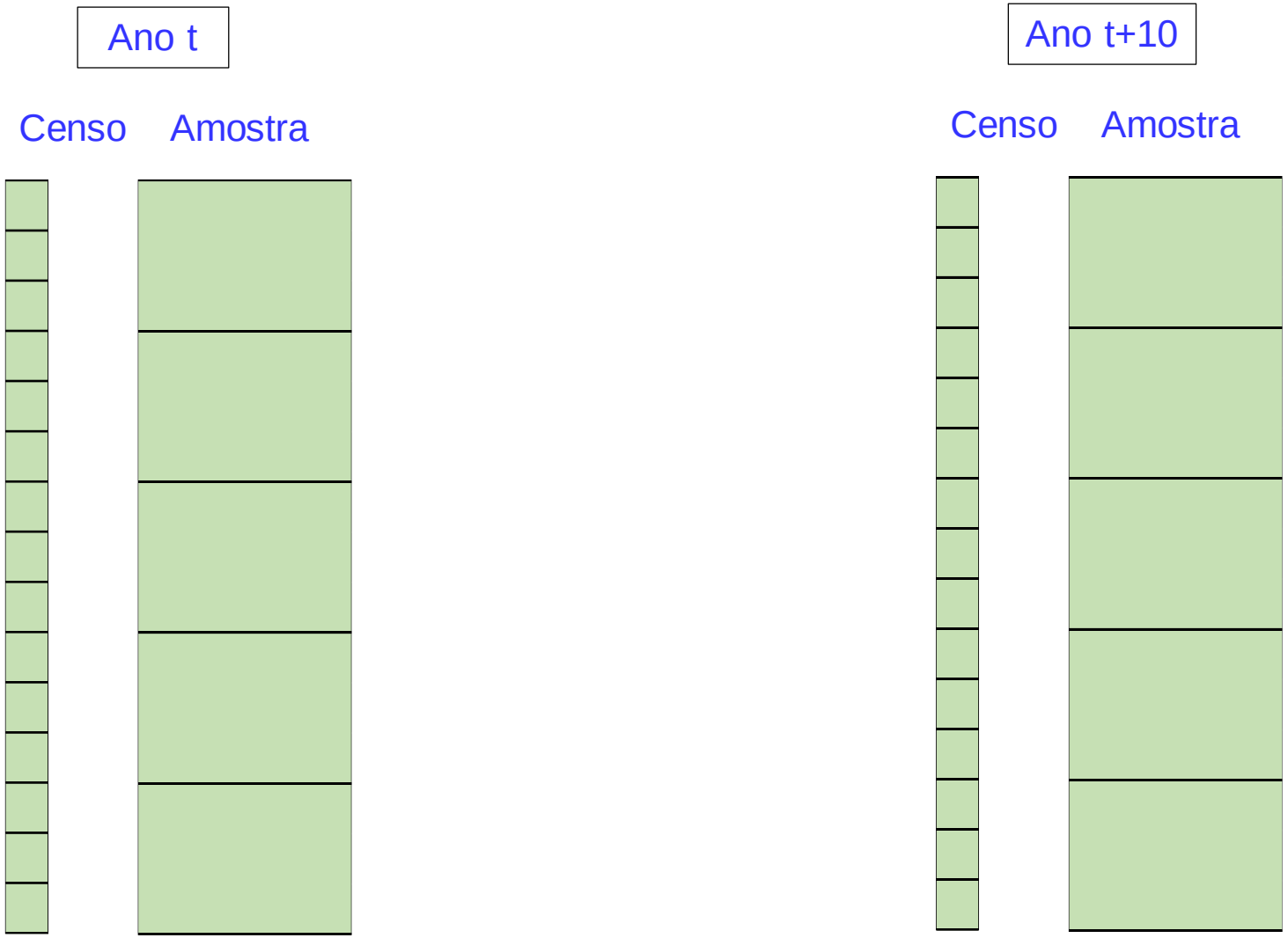
Nos **municípios pequenos**, corresponde ao **município** inteiro.

Nos **municípios maiores**, formada por **agrupamentos de setores censitários contíguos**.

O **menor tamanho** de uma área de ponderação não municipal é de **400 domicílios particulares ocupados na amostra**.

Tamanhos das áreas de ponderação variam bastante (tanto em população como em território).

Censo Demográfico Decenal



Amostragem no Censo

Amostras de domicílios selecionadas de forma independente em cada **setor censitário**.

Em cada setor censitário, **sorteio de domicílios por** método que aproxima **amostragem aleatória simples sem reposição**.

Em cada domicílio selecionado, **informações** levantadas **sobre todos os moradores**.

Amostragem no Censo

Plano amostral:

- **Amostragem estratificada simples** de domicílios;
  - **Estrato** = setor censitário;
- **Amostragem estratificada e conglomerada simples** de moradores
  - **Estrato** = setor censitário;
  - **Conglomerado** = domicílio.

Amostragem no Censo

Em 2010, **fração amostral** empregada para seleção de domicílios variou conforme o **tamanho do município**.

Tabela 11.2 - Fração amostral e número de municípios, segundo as classes de tamanho da população dos municípios - Censo Demográfico 2010

Classes de tamanho da população dos municípios (habitantes)	Fração amostral de domicílios	Número de municípios
Total	**11%**	**(1) 5 565**
Até 2 500	50%	260
Mais de 2 500 até 8 000	33%	1 912
Mais de 8 000 até 20 000	20%	1 749
Mais de 20 000 até 500 000	10%	1 604
Mais de 500 000	5%	40

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais e Coordenação de Métodos e Qualidade.

Nota: Cálculo com base nas estimativas de população residente para 1º de julho de 2009.

(1) Inclui o Distrito Estadual de Fernando de Noronha e o Distrito Federal.

Amostragem no Censo

Dados são 'complexos' porque a amostragem praticada para obtê-los envolve:

- **Estratificação**,
- **Conglomeração** (no caso dos dados de pessoas),
- **Probabilidades desiguais de seleção** (por município),
- **Observações com pesos desiguais** (devido à **calibração** dos pesos).

Consequência do uso de amostragem: dados da amostra precisam ser ponderados para produzir estimativas adequadas para a população.

Dados individuais na amostra do Censo

Curso6.R | Curso1.R | domicrj_dat | Curso6.rmd

Filter

	v0001	v0002	v0011	v0300	v0010	v1001	v1002	v1003	v
1	33	00100	3300100003001	00012833	10.720090	3	05	013	(
2	33	00100	3300100003001	00020358	19.010606	3	05	013	(
3	33	00100	3300100003001	00027895	8.929838	3	05	013	(
4	33	00100	3300100003001	00042944	14.050406	3	05	013	(
5	33	00100	3300100003001	00052534	12.335628	3	05	013	(
6	33	00100	3300100003001	00053400	19.041126	3	05	013	(
7	33	00100	3300100003001	00072097	9.197980	3	05	013	(
8	33	00100	3300100003001	00072357	10.874790	3	05	013	(
9	33	00100	3300100003001	00081428	10.313049	3	05	013	(
10	33	00100	3300100003001	00091807	13.713995	3	05	013	(
11	33	00100	3300100003001	00109266	9.549809	3	05	013	(
12	33	00100	3300100003001	00109629	8.999881	3	05	013	(
13	33	00100	3300100003001	00127957	14.559630	3	05	013	(
14	33	00100	3300100003001	00130577	7.789386	3	05	013	(
15	33	00100	3300100003001	00133609	14.531634	3	05	013	(

Dados individuais na amostra do Censo

Quadro 1 – Registro típico de domicílio

Área de Ponderação (v0011)	Domicílio (v0300)	Peso (v0010)	Y_1	Y_2	...	Y_J

Variáveis Y_1-Y_J usualmente consideradas na análise.

Variáveis de estratificação e peso algumas vezes ignoradas na análise. Isto está OK?

A resposta geralmente é NÃO, como veremos adiante.

Algumas análises ingênuas da amostra de domicílios

```
## Analisa dados da amostra de domicílios do RJ
# Lê dados da amostra de domicílios do RJ
domicrj_dat <- readRDS(file="domicrj_dat.rds")
# Cria e modifica variáveis no arquivo de dados
domicrj_dat <- mutate(domicrj_dat,
                uf = factor(v0001 ,
                            levels = c( 33 ) ,
                            labels = c( "Rio de Janeiro" ) ),
                v0201 = factor(v0201,
                           levels = c(1:6),
       labels = c("Próprio de algum morador - já pago",
                  "Próprio de algum morador - ainda pagando",
                  "Alugado",
                  "Cedido por empregador",
                  "Cedido de outra forma",
                  "Outra condição") ) )
```

Algumas análises ingênuas da amostra de domicílios

```
# Tabula variável 'condição de ocupação do domicílio'
condicao.domicilio <- 100*table(domicrj_dat$v0201) / sum(!is.na
(domicrj_dat$v0201))
condicao.domicilio
```

```
##       Próprio de algum morador - já pago
##                                  "71,82"
## Próprio de algum morador - ainda pagando
##                                  " 3,20"
##                                  Alugado
##                                  "18,37"
##                    Cedido por empregador
##                                  " 1,92"
##                    Cedido de outra forma
##                                  " 3,99"
##                           Outra condição
##                                  " 0,69"
```

Algumas análises ingênuas da amostra de domicílios

```
# Calcula média do aluguel pago em domicílios alugados
media_aluguel <-  mean(domicrj_dat$v2011, na.rm=TRUE)
fprop(media_aluguel)

## [1] "447,50"

# Calcula média da renda domiciliar per capita
media_rdpc <- mean(domicrj_dat$v6531, na.rm=TRUE)
fprop(media_rdpc)

## [1] "1.115,90"

# Estima média da renda domiciliar per capita em domicílios
alugados
domalugrj_dat <-  subset(domicrj_dat, v0201=="Alugado")
media_rdpc_domalug <- mean(domalugrj_dat$v6531, na.rm=TRUE)
fprop(media_rdpc_domalug)

## [1] "1.248,80"
```

Algumas análises ingênuas da amostra de domicílios

```
# Conta número de domicílios da amostra total do RJ
total_domicilios <- table(domicrj_dat$uf)
ftot(total_domicilios)

## Rio de Janeiro
##      "370.244"

# Tabula número de domicílios da amostra por área de ponderação
totais_areapond <-  table(domicrj_dat$v0011)
# Calcula total de moradores em domicílios (população residente
)
# do estado do Rio de Janeiro
total_pessoas <-  sum(domicrj_dat$v0401)
ftot(total_pessoas)

## [1] "1.143.650"
```

Estes números fazem sentido?

O Problema

Dados da amostra, sozinhos, não 'representam' a população.

Por exemplo, considere o total populacional. Ele pode ser particionado em duas componentes:

$$Y = \sum_{i \in U} y_i = \sum_{i \in s} y_i + \sum_{i \in (U-s)} y_i$$

Nessa decomposição fica evidente que a soma da parte observada na amostra sempre vai 'subestimar' o total da população (as unidades na parcela U-s não será observada).

Como resolver: usando teoria da amostragem.

A Solução

Uma **amostra** $s=\{i_1, i_2, ..., i_n\}$ é qualquer **subconjunto** não vazio de unidades da população U ($s \subset U$) selecionadas para observação visando estimar os parâmetros de interesse.

Uma amostra de **tamanho** n é uma amostra contendo n **unidades** selecionadas da população U.

$i \in s$ designa que a unidade i foi incluída na amostra.

Dados amostrais → $y_{i_1}, y_{i_2}, \ldots, y_{i_n}$

Total (soma) amostral: $t(s) = t = \sum_{i \in s} y_i$

A Solução

Considere que o objetivo principal é usar os **dados amostrais** $y_{i_1}, y_{i_2}, \ldots, y_{i_n}$ para **estimar** $Y = \sum_{i \in U} y_i$.

Um objetivo secundário é conseguir medir / estimar também a **precisão** / **margem de erro** **da estimativa** produzida para Y.

Um estimador para o total é dado por:

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} \pi_i^{-1} y_i \Rightarrow \text{Estimador de } \textbf{Horvitz-Thompson}$$

Este estimador está definido para qualquer plano amostral ou variável, desde que $\pi_i > 0 \ \forall i \in U$.

Propriedades do estimador

$E_p\left(\hat{Y}_{HT}\right) = Y$ ➔ estimador HT é não enviesado.

$$V_p\left(\hat{Y}_{HT}\right) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \left(\frac{y_i}{\pi_i} \frac{y_j}{\pi_j}\right)$$

$$\hat{V}\left(\hat{Y}_{HT}\right) = \sum_{i \in s} \sum_{j \in s} \frac{\left(\pi_{ij} - \pi_i \pi_j\right)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} \frac{y_j}{\pi_j}\right)$$

Para amostras grandes, distribuição aproximada do estimador é Normal. Logo

$IC(Y;1\text{-}\alpha) = \hat{Y}_{HT} \mp z_{\alpha/2} \left[V_p\left(\hat{Y}_{HT}\right)\right]^{1/2}$ é um IC de nível 1-α.

Estimação da média populacional

$$\hat{\bar{Y}}_{HT} = \left(\sum_{i \in s} \pi_i^{-1} y_i\right) \Big/ N \qquad \text{Horvitz-Thompson}$$

$$\hat{\bar{Y}}_{H} = \left(\sum_{i \in s} \pi_i^{-1} y_i\right) \Big/ \left(\sum_{i \in s} \pi_i^{-1}\right) \qquad \text{Hàjek}$$

O primeiro estimador só é viável se N (tamanho da população) for conhecido.

No segundo estimador, o denominador ($\sum_{i \in s} \pi_i^{-1}$) fornece uma estimativa do tamanho da população.

Para detalhes sobre estimadores, consultar, por exemplo:

Bolfarine e Bussab (2005).

Revisitando análises da amostra de domicílios

```
## Operação para criar objeto do plano amostral com dados de
domicílios
# Calcula tamanhos da população em cada área de ponderação
tamanho_pop <- aggregate(v0010 ~ v0011, data=domicrj_dat, FUN="sum")
# Ajusta nomes das colunas do arquivo com tamanhos populacionais
names(tamanho_pop) <- c("v0011", "Ndompop")
# Agrega variável com tamanhos populacionais ao arquivo de dados
domicrj_dat <- inner_join(domicrj_dat, tamanho_pop, by="v0011")
# Adiciona estrutura do plano amostral aos dados da amostra
domicrj_plan <- svydesign(data=domicrj_dat,
                          ids = ~1,
                          strata = ~v0011,
                          fpc = ~Ndompop ,
                          weights = ~v0010)
# Armazena dados de domicílios do RJ num arquivo permanente
saveRDS(domicrj_plan, file="domicrj_plan.rds")
```

Revisitando análises da amostra de domicílios

```
## Análises dos dados da amostra de domicílios incorporando plano
amostral
# Estima número de domicílios total do RJ
total_est_domicilios <-  svytable(~ uf, domicrj_plan)
ftot(total_est_domicilios)

## uf
## Rio de Janeiro
##    "5.299.014" ➔ Compare com 370.244 (obtido só na amostra)

# Estima número de domicílios por área de ponderação
total_est_domic_areapond <- svytable(~ v0011, domicrj_plan)
# Estima total de moradores (população residente) em domicílios
# do estado do Rio de Janeiro
total_est_pessoas <-  svytotal(~ v0401, domicrj_plan)
ftot(total_est_pessoas)

##        v0401
## "15.989.929"  ➔ Compare com 1.143.650 (obtido acima)
```

Revisitando análises da amostra de domicílios

```
# Tabula variável 'condição de ocupação do domicílio'
condicao.domicilio <-  svytable( ~ v0201 , domicrj_plan, Ntotal = 100)
fprop(condicao.domicilio)
```

```
## v0201
##       Próprio de algum morador - já pago
##                                    "71,73"
## Próprio de algum morador - ainda pagando
##                                    " 3,53"
##                                   Alugado
##                                    "18,91"
##                     Cedido por empregador
##                                    " 1,46"
##                     Cedido de outra forma
##                                    " 3,68"
##                              Outra condição
##                                    " 0,70"
```

Facilitando a comparação

Condição de Ocupação do Domicílio	Estimativa simples (%)	Estimativa com peso (%)
Próprio de algum morador - já pago	71,82	71,73
Próprio de algum morador - ainda pagando	3,20	3,53
Alugado	18,37	18,91
Cedido por empregador	1,92	1,46
Cedido de outra forma	3,99	3,68
Outra condição	0,69	0,70
Total	99,99	100,01

Revisitando análises da amostra de domicílios

```
# Estima média do aluguel pago em domicílios alugados
media_est_aluguel <- svymean(~ v2011, domicrj_plan, na.rm=TRUE)
fprop(media_est_aluguel)
##    v2011
## "478,18"    ➔ Compare com 447,50 (média sem pesos)
# Estima média da renda domiciliar per capita
media_est_rdpc <- svymean(~ v6531, domicrj_plan, na.rm=TRUE)
fprop(media_est_rdpc)
##      v6531
## "1.231,14"  ➔ Compare com 1.115,90 (média sem pesos)
# Estima média da renda domiciliar per capita em domicílios alugados
media_est_rdpc_domalug <- svymean( ~ v6531 ,
                                   subset(domicrj_plan, v0201==
"Alugado"), na.rm=TRUE )
fprop(media_est_rdpc_domalug)
##      v6531
## "1.374,00"  ➔ Compare com 1.248,80 (média sem pesos)
```

Resumo até aqui

Podemos **ignorar o plano amostral** e usar dados da amostra do Censo 'sem pesos'?

➔ A resposta é **NÃO**.

Em geral, podem ser verificadas diferenças em:

- **Estimativas pontuais**;
- Estimativas de variância;
- Intervalos de confiança;
- Distribuições de estatísticas de teste;
- Graus de liberdade.

Resumo até aqui

Como podemos detectar / medir essas diferenças?

R: Uma forma simples é considerar **duas análises**:

- Uma ignorando os pesos e o plano amostral;
- Outra considerando esses aspectos;

e então comparar os dois resultados.

Mas a **recomendação** é analisar os dados considerando os pesos amostrais.

Análises sem considerar os pesos levam a **estimativas enviesadas** dos parâmetros de interesse!

Resumo até aqui

Até aqui, focamos mais na estimação pontual de parâmetros de interesse.

Mas em como dados são de amostra, é importante também estimar a **precisão das estimativas** (**erro padrão**; **intervalo de confiança**).

Se não podemos ignorar o plano amostral nas análises, como ajustar os procedimentos para fazer inferência usando dados amostrais?

R: Usando **ferramentas de estimação** adequadas.

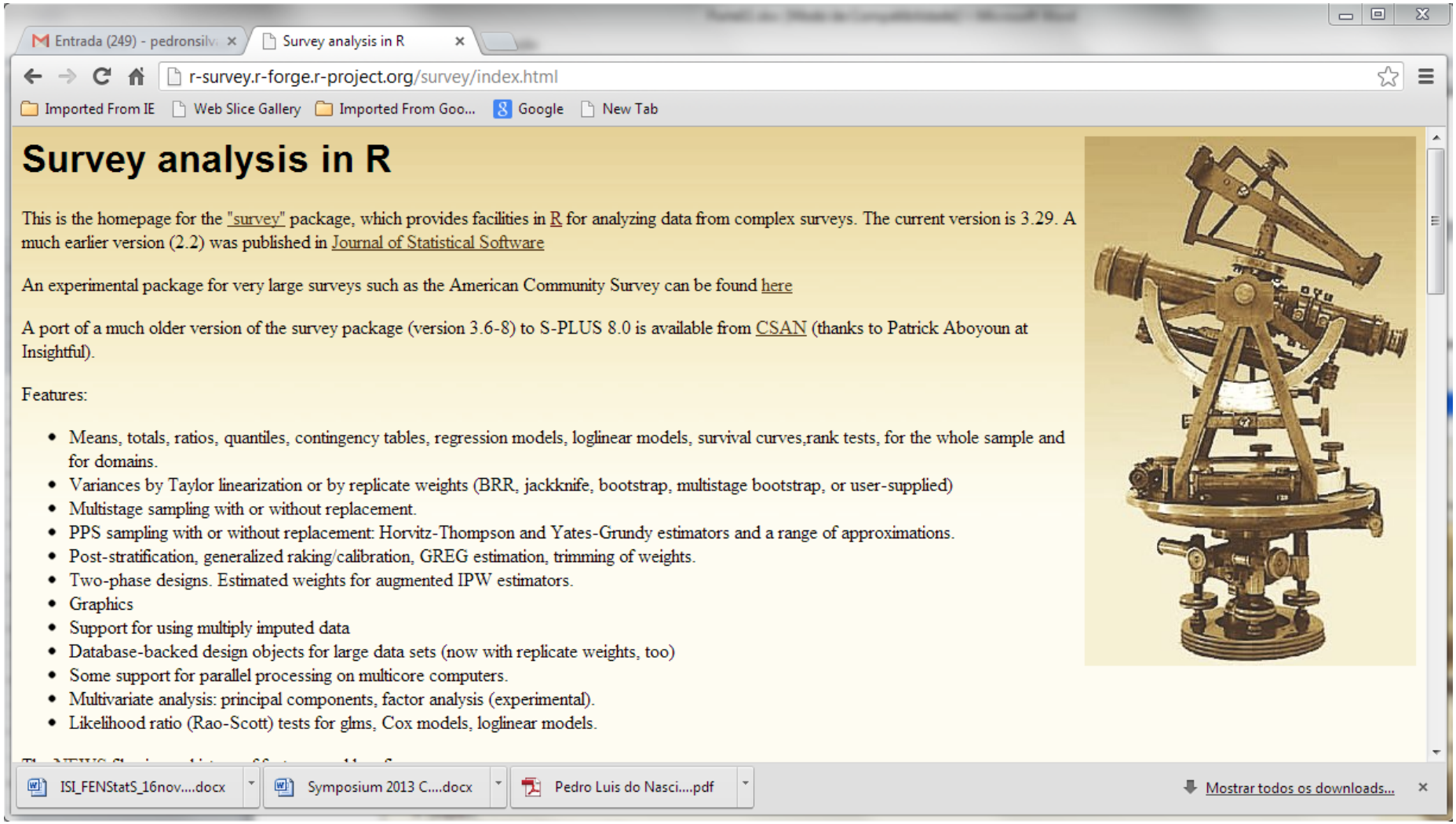
O Pacote *survey* do R

Versão corrente é a 3.32-1.

Pacote ('*library*') elaborado e mantido por Thomas Lumley, da Universidade de Auckland (Nova Zelândia).

Livro publicado (Lumley, 2011) pelo autor apresenta:

- **Teoria** 'clássica' para análise de dados amostrais complexos;
- **Recursos** do pacote *survey* para análise de dados amostrais;
- Inúmeros **exemplos** com dados reais.


Entrada (249) - pedronsilv
Survey analysis in R
r-survey.r-forge.r-project.org/survey/index.html
Imported From IE
Web Slice Gallery
Imported From Goo...
Google
New Tab
Survey analysis in R
This is the homepage for the "survey" package, which provides facilities in R for analyzing data from complex surveys. The current version is 3.29. A much earlier version (2.2) was published in Journal of Statistical Software
An experimental package for very large surveys such as the American Community Survey can be found here
A port of a much older version of the survey package (version 3.6-8) to S-PLUS 8.0 is available from CSAN (thanks to Patrick Aboyoun at Insightful).
Features:
Means, totals, ratios, quantiles, contingency tables, regression models, loglinear models, survival curves,rank tests, for the whole sample and for domains.
Variances by Taylor linearization or by replicate weights (BRR, jackknife, bootstrap, multistage bootstrap, or user-supplied)
Multistage sampling with or without replacement.
PPS sampling with or without replacement: Horvitz-Thompson and Yates-Grundy estimators and a range of approximations.
Post-stratification, generalized raking/calibration, GREG estimation, trimming of weights.
Two-phase designs. Estimated weights for augmented IPW estimators.
Graphics
Support for using multiply imputed data
Database-backed design objects for large data sets (now with replicate weights, too)
Some support for parallel processing on multicore computers.
Multivariate analysis: principal components, factor analysis (experimental).
Likelihood ratio (Rao-Scott) tests for glms, Cox models, loglinear models.
ISI_FENStatS_16nov....docx
Symposium 2013 C....docx
Pedro Luis do Nasci....pdf
Mostrar todos os downloads...

Princípios condutores do desenho do pacote *survey*

- **Facilidade de manutenção** e depuração mediante reutilização de código.
- **Velocidade e memória não são prioridade**: só otimiza rotinas quando há um 'caso real de uso' demandando solução.
- **Rápida liberação de novas versões**, de modo que erros e outras infelicidades sejam descobertas e reparadas.

Usos propostos

- **Análise secundária de dados** de pesquisas nacionais (**R** é familiar a estatísticos não ligados à área de amostragem).
- **Preparação dos dados** das amostras para análise / disseminação (pelos estatísticos das agências produtoras).
- **Pesquisa em métodos** (devido às características de programação do **R**).
- **Ensino** (facilita integração com ensino de outros métodos estatísticos, onde **R** também é usado).

Características e funcionalidades

Descrição de planos amostrais: *svydesign*().

Calibração e **pós-estratificação**: *calibrate*(), *poststratify()*.

Estatísticas descritivas: médias, totais, quantis, razões, etc. - *svymean*(), *svytotal*(), *svyquantile*(), *svyratio*(), etc.

Estimação para domínios: *subset*(), *svyby*().

Tabelas de contingência: *svytable*(), *svychisq*(), *svyloglin*().

Gráficos: histogramas *svyhist*(), diagramas de dispersão *svyplot*(), suavizadores *svysmooth*().

Modelos de regressão: *svyglm*(), *svyloglin*(), *svyolr*().

Métodos usados para estimação

Unidades são amostradas com **probabilidades de inclusão** π_i **conhecidas** de uma população de tamanho N, para obter uma amostra de tamanho n.

O problema 'usual' de inferência considerando o plano amostral é **estimar quantidades populacionais** definidas caso toda a população fosse observada.

Estimação

A estimação de um **total populacional** é simples.

Um estimador não viciado do total $Y = \sum_{i \in U} y_i$ é dado por:

$$\hat{Y}_{HT} = \sum_{i \in s} \pi_i^{-1} y_i$$

Outros estimadores usuais de total são da forma

$$\hat{Y}_C = \sum_{i \in s} (\pi_i^{-1} g_i) y_i = \sum_{i \in s} w_i y_i$$

com g_i = **fator de calibração** do peso amostral básico π_i^{-1}, tal que $\sum_{i \in s} (\pi_i^{-1} g_i) \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i$.

Equações de estimação

A estimação de **outros parâmetros** (p. ex. médias e razões) segue da estimação de totais.

Se uma **quantidade populacional** de interesse θ é solução da **equação de estimação**:

$$\sum_{i \in U} u_i(\theta) = 0$$

Então um **estimador amostral** $\hat{\theta}$ vai ser a solução de

$$\sum_{i \in s} w_i u_i(\theta) = 0$$

com $w_i = \pi_i^{-1}$ ou $w_i = \pi_i^{-1} g_i$ ou outro peso adequado.

Exemplo 1: Estimação da Média Populacional

Defina $u_i(\theta) = y_i - \theta \ \forall i \in U$. Então:

$$\sum_{i \in U} u_i(\theta) = 0 \ \Leftrightarrow \ \sum_{i \in U} (y_i - \theta) = 0 \ \Leftrightarrow \ \theta = \sum_{i \in U} y_i / N = \bar{Y}.$$

Consequentemente, o estimador amostral para a **média populacional** é obtido resolvendo:

$$\sum_{i \in s} w_i (y_i - \theta) = 0 \ \Leftrightarrow \ \sum_{i \in s} w_i y_i = \theta \sum_{i \in s} w_i \Rightarrow$$

Logo:

$$\hat{\theta} = \sum_{i \in s} w_i y_i / \sum_{i \in s} w_i \ .$$

Exemplo 2: Estimação de uma razão de totais

Defina $u_i(\theta) = y_i - \theta x_i \ \forall\, i \in U$. Então:

$$\sum_{i \in U} (y_i - \theta x_i) = 0 \Leftrightarrow \sum_{i \in U} y_i = \theta \sum_{i \in U} x_i \Rightarrow$$

$$\theta = \sum_{i \in U} y_i / \sum_{i \in U} x_i$$

Consequentemente, o estimador amostral para a **razão de totais** é obtido resolvendo:

$$\sum_{i \in s} w_i (y_i - \theta x_i) = 0 \Leftrightarrow \sum_{i \in s} w_i y_i = \theta \sum_{i \in s} w_i x_i \Rightarrow$$

Logo:

$$\hat{\theta} = \sum_{i \in s} w_i y_i / \sum_{i \in s} w_i x_i \quad .$$

Estimação

Estimação da **precisão** (**erros padrão**) segue expressões disponíveis para muitos planos amostrais descritos na literatura.

Para muitos casos, **erros padrão** são obtidos mediante aproximações:

- **Linearização de Taylor** (método delta);
- **Método do Conglomerado Primário;**
- **Métodos de reamostragem**.

Para detalhes, consultar por exemplo Pessoa & Silva (1998) e Lumley (2011).

Estratégia do pacote *survey*

Coleções de informações relacionadas armazenadas juntas num único **objeto** que contém:

- **Dados**;
- **Metadados** relevantes (rótulos de categorias, etc.);
- Informações sobre a **estrutura do plano amostral** usado para obter os dados;
- Informações sobre **método(s) usado(s) para estimação de erros padrões**.

Dados de entrada têm que ser armazenados num '***data frame***'.

Descrevendo um plano amostral no *survey*

A função ***svydesign()*** é a que permite descrever a **estrutura de um plano amostral** para o pacote ***survey***.

Possui recursos para especificar:

- **Estratificação**;
- **Conglomeração**;
- Observações com **pesos desiguais**, para lidar com **probabilidades desiguais** de seleção; e
- **Métodos** a serem empregados para estimar **erro padrão**.

Depois de aplicada, os **metadados** sobre o plano amostral são **armazenados junto** dos dados da pesquisa num objeto especial (lista) reconhecido pelas demais funções do pacote.

Etapas necessárias para usar pacote *survey*

1. Especificar a **estrutura do plano amostral** usado para obter os dados que quer analisar ➔ função *svydesign*().
2. Se aplicável, efetuar a calibração dos pesos e criar o objeto adequado para as análises ➔ funções *calibrate*(), *rake*() ou *poststratify()*.
3. Especificar **análise de interesse** – por exemplo, função que permite estimar médias ou totais populacionais ➔ funções *svymean*() e *svytotal*().

Plano amostral para amostra de domicílios do Censo

```
## Operação para criar objeto do plano amostral com dados de
domicílios
# Calcula tamanhos da população em cada área de ponderação
tamanho_pop <- aggregate(v0010 ~ v0011, data=domicrj_dat, FUN="sum")
# Ajusta nomes das colunas do arquivo com tamanhos populacionais
names(tamanho_pop) <- c("v0011", "Ndompop")
# Agrega variável com tamanhos populacionais ao arquivo de dados
domicrj_dat <- inner_join(domicrj_dat, tamanho_pop, by="v0011")
# Adiciona estrutura do plano amostral aos dados da amostra
domicrj_plan <- svydesign(data=domicrj_dat,
                          ids = ~1,
                          strata = ~v0011,
                          fpc = ~Ndompop )
# Armazena dados de domicílios do RJ num arquivo permanente
saveRDS(domicrj_plan, file="domicrj_plan.rds")
```

Plano amostral para amostra de pessoas do Censo

```
## Operação para criar objeto do plano amostral com dados de
pessoas
# Calcula tamanhos da população em cada área de ponderação
tamanho_pop <- aggregate(v0010 ~ v0011, data=pesrj_dat, FUN="sum")
# Ajusta nomes das colunas do arquivo com tamanhos populacionais
names(tamanho_pop) <- c("v0011", "Npespop")
# Agrega variável com tamanhos populacionais ao arquivo de dados
pesrj_dat <- inner_join(pesrj_dat, tamanho_pop, by="v0011")
# Adiciona estrutura do plano amostral aos dados da amostra
pesrj_plan <- svydesign(data= pesrj_dat,
                        ids = ~v0300,
                        strata = ~v0011,
                        fpc = ~Npespop,
                        weights = ~v0010) )
# Armazena dados de pessoas do RJ num arquivo permanente
saveRDS(pesrj_plan, file="pesrj_plan.rds")
```

Comentários

Especificação da estrutura do plano amostral costuma ser feita **uma única vez para cada pesquisa** ou conjunto de dados.

Análises incorporando plano amostral são tão **simples** de obter quanto análises ignorando o plano amostral.

Análises usando ferramentas do pacote fornecem objetos que podem ser reutilizados para novos cálculos e/ou para exportação dos resultados.

Exemplos de análises descritivas – dados de pessoas

```
# Carrega objeto com dados de domicílios
pesrj_plan <- readRDS("pesrj_plan.rds ")
# Cria e ajusta variáveis
pesrj_plan <-
  update(pesrj_plan,
         sexo = factor( v0601 ,
         labels = c( "masculino" , "feminino" ) ) ,
         UF = factor(v0001 ,
         levels = c( 33 ) ,
         labels = c( "Rio de Janeiro" ) ) )
# Conta número de pessoas da amostra total do RJ
n_pessoas <- svyby( ~ one , ~ UF , pesrj_plan , unwtd.count )
ftot(n_pessoas)
```

```
##              UF    counts   se
## Rio de Janeiro 1.143.650    0
```

Exemplos de análises descritivas – dados de pessoas

```
# Estima total de moradores do estado do Rio de Janeiro
N_pessoas <- svytotal( ~ one , pesrj_plan )
ftot(N_pessoas)
##           one
## "15.989.929"


# Estima total de pessoas por sexo no RJ
totais_sexo <- svyby( ~ one , ~ sexo , pesrj_plan , svytotal )
ftot(totais_sexo)
##       sexo         one        se
## masculino   7.625.679    10.286
##  feminino   8.364.250    10.262
```

Exemplos de análises descritivas – dados de pessoas

```
# Tabula variável condição de ocupação na semana de referência
totais_condocup <- svyby( ~ one , ~ v6910 , pesrj_plan , svytotal )
ftot(totais_condocup)
##             v6910          one       se
## 1         Ocupado   7.151.619   10.935
## 2     Desoculado     663.108    3.614


# Tabula variável condição de atividade na semana de referência
totais_condativ <- svyby( ~ one , ~ v6900 , pesrj_plan , svytotal )
ftot(totais_condativ)
##          v6900          one       se
## 1        Ativo   7.814.727   11.164
## 2     Inativo   6.093.446   11.337
```

Exemplos de análises descritivas – dados de pessoas

```
# Estima taxa de desocupação
taxa_desocup <- svyratio( ~ (v6910==2) ,  ~ (v6900==1),  pesrj_
plan, na.rm=TRUE )
fprop(100*coef(taxa_desocup))
## v6910 == 2/v6900 == 1
##                 "8,49"
fprop(100*SE(taxa_desocup))
## v6910 == 2/v6900 == 1
##                 "0,05"
```

Exemplos de análises descritivas

```
# Estima média da renda domiciliar per capita
media_est_rdpc <- svymean(~ v6531, domicrj_plan, na.rm=TRUE)
fprop(coef(media_est_rdpc))

##      v6531
## "1.231,14"

fprop(SE(media_est_rdpc))

##       v6531
## v6531 "6,85"

fprop(confint(media_est_rdpc))

##       2.5 %      97.5 %
## v6531 "1.217,71" "1.244,57"
```

Exemplos de análises descritivas

```
# Estima média da RDPC por área de ponderação
media_est_rdpc_apond <- svyby(~ v2011, ~v0011, domicrj_plan, sv
ymean, na.rm=TRUE)
result_apond <- cbind(coef(media_est_rdpc_apond)[1:10],
                      SE(media_est_rdpc_apond)[1:10],
                      confint(media_est_rdpc_apond)[1:10,])
colnames(result_apond) <- c("Media_rdpc", "SE_Media_rdpc", "LI_
Media_rdpc", "LS_Media_rdpc")
fprop(result_apond)
```

Exemplos de análises descritivas

```
##                Media_rdpc SE_Media_rdpc LI_Media_rdpc LS_Media_rdpc
## 3300100003001 "410,57"    " 18,21"      "374,87"      "446,26"
## 3300100003002 "328,36"    " 10,25"      "308,28"      "348,45"
## 3300100003003 "502,58"    " 21,75"      "459,95"      "545,21"
## 3300100003004 "453,65"    " 20,48"      "413,51"      "493,79"
## 3300100003005 "452,41"    " 41,63"      "370,81"      "534,00"
## 3300100003006 "488,95"    " 29,88"      "430,38"      "547,52"
## 3300100003007 "327,91"    " 11,96"      "304,48"      "351,35"
## 3300159001001 "250,87"    "  9,01"      "233,22"      "268,53"
## 3300209003001 "221,76"    " 10,81"      "200,58"      "242,94"
## 3300209003002 "370,43"    " 27,13"      "317,25"      "423,60"
```

Alguns sites úteis

https://djalmapessoa.github.io/adac/

http://asdfree.com/

Resumindo

1. Considere pesos das observações ao calcular estimativas pontuais.
2. Considere a estrutura do plano amostral (estratificação, conglomeração e pesos) ao calcular estimativas de variância e ao ajustar modelos com dados da amostra.
3. Respeite os limites da 'geografia da amostra'.
4. Procure conhecer bem a metodologia da pesquisa cujos dados vai usar para analisar.

Referências

Bolfarine, H., & Bussab, W. de O. (2005). *Elementos de Amostragem*. *Projeto Fisher*. São Paulo: Editora Edgard Blücher.

IBGE. (2016). *Metodologia do Censo Demográfico 2010, 2a edição*. Rio de Janeiro, Brasil: Instituto Brasileiro de Geografia e Estatística. Disponível em http://biblioteca.ibge.gov.br/visualizacao/livros/liv95987.pdf.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley Series in Survey Methodology. Hoboken: John Wiley & Sons.

Pessoa, D. G. C., & Silva, P. L. do N. (1998). *Análise de dados amostrais complexos*. São Paulo: Associação Brasileira de Estatística.

SILVA, P. L. d. N. (2004). *Calibration Estimation: When and Why, How Much and How*. Rio de Janeiro: IBGE, Textos para Discussão da Diretoria de Pesquisas, número 15.