

# Aula 0: Estimação com a PNAD Contínua

## Conceitos básicos e prática usando o R

Guilherme Jacob

02/08/2021

## 1 Conceitos

- 1 Amostragem
- 2 Teoria do Erro Total
- 3 O que é a PNAD Contínua?
- 4 Plano amostral da PNAD Contínua
- 5 Mensuração do rendimento na PNAD Contínua
- 6 Leitura sugerida

## 2 Prática com o R

- 1 Pacote survey
- 2 Plano amostral da PNAD Contínua com svydesign
- 3 Estimando médias com a função svymean
- 4 Estimação para domínios
- 5 Estimativas para vários domínios
- 6 Principais recomendações

# Parte 1

## Conceitos

# Amostragem

- Uma população finita  $U$  de  $N$  indivíduos, identificados pelos rótulos 1, 2, ...,  $N$ .
  - $U = \{1, 2, \dots, N\} = \{i\}_{i=1}^N$
- Se dispomos de recursos ilimitados, podemos calcular diversas quantidades nesta população;
- Denotando o valor da variável  $y$  para o indivíduo  $i \in U$  por  $y_i$ , podemos calcular, por exemplo:
  - O total de  $y$  na população:  $Y = \sum_{i=1}^N y_i$ ;
  - A média populacional:  $\bar{Y} = \sum_{i=1}^N y_i / N$ ;
  - O índice de Gini de  $y$ , etc.

# Amostragem

- Porém, não podemos coletar informação sobre  $N$  indivíduos.
- Portanto, extraímos uma amostra  $S$  de  $n$  indivíduos de acordo com um plano amostral:
  - Conhecemos as probabilidades de seleção  $\pi_i > 0, \forall i \in U$ ;
  - Idealmente, também conhecemos as probabilidades conjuntas de seleção  $\pi_{ij} > 0, \forall i, j \in U$ .
  - Sorteamos os indivíduos de acordo com estas probabilidades.
- Planos amostrais probabilísticos: AAS, Estratificada, Conglomerados, Multi-estágios, etc.

# Amostragem

- Para inferir sobre o parâmetro  $\theta$  da população finita a partir da amostra probabilística  $S$ , utilizamos um estimador  $\hat{\theta}$ ;
  - Estratégia de estimação: plano amostral + estimador.
- Por não conhecermos os valores para todos os  $N$  indivíduos, nossas estimativas têm erros;
  - $EQM(\hat{\theta}) = Var(\hat{\theta}) + B(\hat{\theta})^2$
- Se o nosso estimador  $\hat{\theta}$  é não-viesado,  $B(\hat{\theta}) = 0$ .
  - Logo,  $EQM(\theta) = Var(\hat{\theta})$
- Estimador de Horvitz-Thompson:
  - Amostragem sem reposição;
  - Não-viesado se  $\pi_i > 0, \forall i \in U$ ;
  - Variância estimável se  $\pi_{ij} > 0, \forall i, j \in U$ .

# Amostragem

Pontos principais:

- Importância da estratégia: plano amostral + estimador;
- Sob condições de regularidade, funções de estimadores de HT são:
  - *Assintoticamente* não-viesadas;
  - Variância assintótica pode ser estimada.
- A palavra “modelo” não foi mencionada.
  - Inferência baseada no plano amostral (*design-based inference*)

# Amostragem

## Tipos de parâmetros em pesquisas amostrais

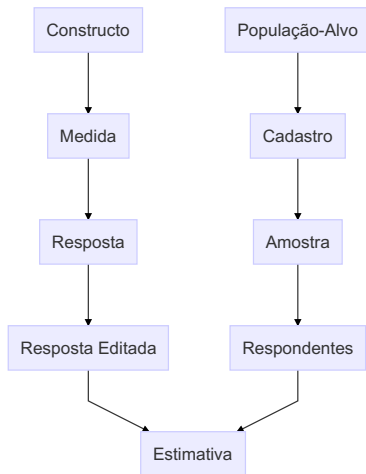
- Descritivos: quantidades na população finita.
  - Prevalência de doenças, taxa de pobreza, média de salários, total de domicílios, medidas de desigualdade, etc.
- Analíticos: parâmetros relativos a causas e associações.
  - Coeficientes em modelos de regressão, correlações entre variáveis, etc.



# Teoria do Erro Total

- A teoria do erro total separa os erros em dois eixos:
  - Mensuração
  - Representação
- Mensuração: erros na informação sobre uma unidade.
- Representação: erros na representação de uma população.

# Teoria do Erro Total



# O que é a PNAD Contínua? (IBGE, 2014)

- Pesquisa Nacional por Amostra de Domicílios Contínua
- População-Alvo:
  - Todas as pessoas moradoras em domicílios particulares permanentes da área de abrangência da pesquisa.
- Abrangência geográfica:
  - Todo o território nacional, excluídas áreas com características especiais.

# Plano Amostral da PNAD Contínua

- Conglomerado em Dois Estágios com Estratificação das Unidades Primárias de Amostragem.

*O que isso significa?*

# Plano Amostral da PNAD Contínua

## *Estratificação*

- Estratificação: sorteio em cada estrato.
- Estratos: agrupamentos baseados em
  - Dependências administrativas;
  - Características sociodemográficas.
- Finalidade:
  - Operacional: Garantir amostras para determinados domínios (UFs, capitais, regiões metropolitanas, etc.)
  - Estatística: melhorar a precisão dos estimadores

# Plano Amostral da PNAD Contínua

## *Conglomerado em Dois estágios*

- Dois estágios de seleção:
  - Sorteio de unidades primárias (UPAs)
  - Sorteio de unidades secundárias (USAs)
- PNADC:
  - UPAs são setores ou grupos de setores censitários;
    - Setores com menos de 60 domicílios particulares permanentes foram combinados até atingirem o tamanho mínimo.
  - USAs são domicílios particulares permanentes ocupados;
- UPAs selecionadas com probabilidade proporcional ao tamanho;
  - Tamanho: número de domicílios particulares permanentes.
- Em cada UPA selecionada, são sorteados 14 USAs por amostragem aleatória simples;
  - Usando o CNEFE.

# Plano Amostral da PNAD Contínua

*Por que isso é importante?*

- A inferência estatística não se baseia em modelos, mas no plano amostral;
- A distribuição amostral depende da probabilidade de seleção de cada unidade na população;
- As hipóteses usuais (independência, por exemplo) não são satisfeitas;
- É assim que (praticamente) todos os institutos de estatística operam.

# Plano Amostral da PNAD Contínua

## *Painel Rotativo*

- Os domicílios são visitados 5 vezes a cada três meses;
- Em cada trimestre, a amostra é composta de 5 grupos:
  - Domicílios que estão na 1ª, 2ª, 3ª, 4ª e 5ª visita.
- À cada trimestre, temos:
  - Domicílios entrando: 1ª visita;
  - Domicílios saindo: 5ª visita;
  - Domicílios de trimestres anteriores que permanecem: 2ª, 3ª e 4ª visitas.



# Mensuração do rendimento na PNAD Contínua

- Informações sobre rendimento são problemáticas:
  - As pessoas podem não querer responder;
  - Esquecer o rendimento de algum morador;
  - “Sub-reportar” rendimento;
  - Valores incorretos ou suspeitos.

## Mensuração do rendimento na PNAD Contínua

Hoffmann, Botassio e Jesus (2019, p. 256–257) apontam as seguintes perguntas sobre rendimentos efetivamente recebidos na PNADC:

- a. Rendimento bruto/retirada mensal que recebeu/fez, em dinheiro, no trabalho principal;
- b. Rendimento bruto/retirada mensal que recebeu/fez, em produtos ou mercadorias, no trabalho principal;
- c. Rendimento bruto/retirada mensal que recebeu/fez, em dinheiro, no trabalho secundário;
- d. Rendimento bruto/retirada mensal que recebeu/fez, em produtos ou mercadorias, no trabalho secundário;
- e. Rendimento bruto/retirada mensal que recebeu/fez, em dinheiro, em outros trabalhos;
- f. Rendimento bruto/retirada mensal que recebeu/fez, em produtos ou mercadorias, em outros trabalhos;

# Mensuração do rendimento na PNAD Contínua

- g. Rendimento que recebeu de Benefício Assistencial de Prestação Continuada;
- h. Rendimento que recebeu do Programa Bolsa Família;
- i. Rendimento que recebeu de outros programa sociais do governo;
- j. Rendimento que recebeu de aposentadoria ou pensão de instituto de previdência federal (INSS), estadual, municipal, ou do governo federal, estadual, municipal.
- k. Rendimento que recebeu de pensão alimentícia, doação ou mesada em dinheiro de pessoa que não morava no domicílio;
- l. Rendimento que recebeu de aluguel ou arrendamento;
- m. Outros rendimentos não citados: seguro-desemprego, seguro-defeso, bolsa de estudo, juros de caderneta de poupança, etc.

# Mensuração do rendimento na PNAD Contínua

- Rendimento de trabalho costuma ter respostas melhores;
- Mas rendimentos de outras fontes podem ter problemas;
  - Ativos em bancos, aluguéis, etc.
- A PNADC é uma pesquisa sobre força de trabalho;
  - Coleta informações sobre rendimento de trabalho em todas as visitas;
  - Mas rendimentos de outras fontes são investigadas na 1ª e 5ª visitas;
  - Alternativamente, Pesquisa de Orçamentos Familiares (POF).

# Mensuração do rendimento na PNAD Contínua

- A estatística dispõe de técnicas para lidar com esses problemas.
  - Mas elas não fazem milagres e dependem de modelos e suposições.
- Não-resposta de item pode ser atenuada com imputação;
  - Mas isso afeta o cálculo da variância dos estimadores, principalmente quando a taxa de não-resposta é muito alta.

# Leitura sugerida

- IBGE (2014): documentação da PNAD Contínua;
- Hoffmann, Botassio e Jesus (2019) , Capítulo 10: dados de renda na PNAD Contínua e suas limitações;
- Deaton (2019), Capítulo 1: planejamento e estimação com pesquisas domiciliares;
- West, Sakshaug e Kim (2017): impacto da especificação incorreta do plano amostral sobre as estimativas.

## Parte 2

### Prática usando o R

# Pacote survey (Lumley, 2004, 2021)

- Estratégia de estimação: plano amostral + estimador
- Problema: os estimadores mudam de acordo com o plano amostral.
  - Principalmente os estimadores de variância.
- O pacote survey cria um ambiente para aplicar estratégias de estimação.



## Pacote survey (Lumley, 2004, 2021)

- Neste pacote, destacamos duas classes de funções:
  - `svydesign`, que cria objetos que descrevem o plano amostral;
  - Funções de estimação: `svytotal`, `svymean`, `svyquantile`, `svycdf`, `svyglm`, etc.
- Usando funções de estimação com objetos de plano amostral, é possível implementar estratégias de estimação adequadas.

# Plano amostral da PNAD Contínua com svydesign

```
# carrega pacote  
library( survey )  
  
# cria objeto de plano amostral  
pnadc.design <-  
  svydesign( ids = ~ upa + v1008 ,  
            strata = ~ estrato ,  
            weights = ~ v1032 ,  
            data = pnadc.df ,  
            nest = TRUE )
```

# Plano amostral da PNAD Contínua com svydesign

- Cada argumento da função descreve um aspecto do plano amostral;
- Esta função cria um objeto de plano amostral;
  - No nosso caso, o objeto `pnadc.design`.

Esse é o “print” do objeto:

```
## Stratified 2 - level Cluster Sampling design (with replacement)
## With (12087, 150667) clusters.
## svydesign(ids = ~upa + v1008, strata = ~estrato, weights = ~v1032,
##          data = pnadc.df, nest = TRUE)
```

# Estimando médias com a função svymean

- svymean: função que estima médias;
- Variável: rendimento domiciliar per capita def.rdp;
- na.rm = TRUE: tratamento de valores ausentes.

```
svymean( ~def.rdp , pnadc.design , na.rm = TRUE )
```

```
##              mean      SE  
## def.rdp 1406.3 18.599
```

# Estimando médias com a função `svymean`

- Por que `na.rm = TRUE`?
  - NA: valor ausente;
  - Alguns moradores têm valor ausente para o rendimento domiciliar per capita.
  - Por exemplo: empregados que moram no domicílio do empregador.
    - Dupla contagem do rendimento.

# Estimação para domínios

- Às vezes, estamos interessados em domínios (i.e., subpopulações) específicas;
  - Por exemplo, moradores de domicílios na área rural.
- Podemos filtrar essas observações usando a função `subset` sobre o objeto de plano amostral:

```
svymean( ~def.rdp ,
         subset( pnadc.design , v1022 == "Rural" ) ,
         na.rm = TRUE )
```

```
##              mean      SE
## def.rdp 669.15 7.4006
```

# Estimação para domínios

*Posso deletar observações na base de dados em vez de usar `subset` no objeto de plano?*

- Em regra, **não**.
- Motivo: risco de obter estimativas de variância incorretas.
- Em planos amostrais complexos, sempre use a função `subset`:
  - Em regra, ela atribui peso zero para as observações que não nos interessam;
  - Quando possível, ela deleta observações da base para criar um objeto “mais leve”.

# Estimativas para vários domínios

- O método de subset é excelente quando temos um domínio específico.
- Mas isso é um problema quando temos vários domínios:
  - Por exemplo, um comando para cada UF é pouco prático.
- Solução: função `svyby`



# Estimativas para vários domínios

Média do rendimento domiciliar per capita por Grande Região:

```
svyby( ~def.rdpc , ~regiao , pnadc.design ,
       svymean , na.rm = TRUE , deff = TRUE )
```

##	regiao	def.rdpc	se	DEff.def.rdpc
## 1	Norte	871.9841	20.20048	13.134273
## 2	Nordeste	884.3451	19.31939	24.229302
## 3	Sudeste	1720.2881	40.58295	23.413971
## 4	Sul	1701.4766	26.38961	9.698864
## 5	Centro-Oeste	1580.4527	36.24522	11.616859

## Estimativas para vários domínios

*Por que `deff` = TRUE?*

- `DEff` é uma abreviação para *Design Effect* (Kish, 1965);
  - EPA: Efeito do Plano Amostral.
- Não é algo exclusivo da `svyby`, mas de qualquer estimador.
- Ele apresenta o impacto do plano amostral sobre a estimativa da variância do estimador:
  - Por exemplo: um `DEff` = 13 indica que a estimativa correta da variância é 13 vezes maior do que sob AAS.
- Ele pode ser usado para indicar o tamanho do erro em ignorar o plano amostral complexo.

# Principais recomendações

- O plano amostral importa!
- Sempre use a base de dados completa na `svydesign`;
  - Filtre domínios com a função `subset` no objeto de plano amostral.
- Consulte a documentação do pacote `survey`;
- Na dúvida, consulte um estatístico.

## Parte 3

## Referências

## Referências

DEATON, A. **The Analysis of Household Surveys (Reissue Edition with a New Preface): A Microeconometric Approach to Development Policy**. Washington, D.C.: The World Bank, 2019.

HOFFMANN, R.; BOTASSIO, D. C.; JESUS, J. G. DE. **Distribuição de Renda: Medidas de Desigualdade, Pobreza, Concentração, Segregação e Polarização**. 2. ed. São Paulo: Editora da Universidade de São Paulo, 2019.

IBGE. **Pesquisa Nacional por Amostra de Domicílios Contínua: Notas metodológicas**. Rio de Janeiro: IBGE; Instituto Brasileiro de Geografia e Estatística, 2014. Disponível em: <[https://ftp.ibge.gov.br/Trabalho\\_e\\_Rendimento/Pesquisa\\_Nacional\\_por\\_Amostra\\_de\\_Domicilios\\_continua/Notas\\_metodologicas/notas\\_metodologicas.pdf](https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Notas_metodologicas/notas_metodologicas.pdf)>.

KISH, L. **Survey Sampling**. Nova York: John Wiley & Sons, 1965.

# Referências

LUMLEY, T. Analysis of Complex Survey Samples. **Journal of Statistical Software**, v. 9, n. 1, p. 1–19, 2004.

\_\_\_\_. **survey: analysis of complex survey samples**, 2021.

WEST, B. T.; SAKSHAUG, J. W.; KIM, Y. Analytic Error as an Important Component of Total Survey Error. *In*: BIEMER, P. P. *et al.* (Eds.). **Total Survey Error in Practice**. Hoboken, Nova Jersey: John Wiley & Sons, 2017. p. 487–510.