

# Aula 11 - Microdados com R (PNAD Contínua). Pacote PNADcIBGE e Survey

Prof. Dr. José Weligton Félix Gomes

# Pacote PNADcIBGE, survey e dplyr

## 1 – Pacote PNADcIBGE

- Desenvolvido para facilitar o *download*, **importação** e **análise dos dados amostrais** da Pesquisa Nacional por Amostra de Domicílios Contínua realizada pelo Instituto Brasileiro de Geografia e Estatística - IBGE.
- A PNAD Contínua possui **dois tipos de microdados**:
  - **Trimestral**, que contém a parte básica investigada pela pesquisa, contendo variáveis conjunturais de mercado de trabalho referentes a um trimestre civil;
  - **Anual**, que contém temas estruturais específicos investigados na pesquisa para um ano civil.
- Maiores informações sobre a pesquisa e os temas investigados podem ser encontrados no [site oficial do IBGE](#).

# Pacote PNADcIBGE, survey e dplyr

## 2 - Pacote survey

- Devido ao plano amostral utilizado na PNAD Contínua, é necessário que sejam utilizadas ferramentas específicas para a análise de dados amostrais complexos.
- O pacote [survey](#) é um pacote criado especificamente para análise e modelagem de dados provenientes de pesquisas com estes tipos de **planos amostrais**.
- Cálculo de médias, totais, razões, quantis, tabelas de contingência.
- Modelos de regressão, modelos loglineares, curvas de sobrevivência, testes de classificação, para toda a amostra e para domínios (sub - populações).

# Pacote PNADcIBGE, survey e dplyr

Todas essas funções especiais começam com svy e incluem:

- `svytable()` – Realiza uma tabulação cruzada ponderada
- `svymean()` – Exibe a média
- `svyratio()` - Proporção entre as variáveis
- `svychisq()` – Teste Qui-Quadrado (Wald)
- `svytotal()` – Exibe o total ponderado e o erro padrão da variável
- `svyhist()` – Histograma
- `svyplot()` – Gráfico
- `svyglm()` – Regressão
- `svyloglin()` – Regressão

# Pacote PNADcIBGE, survey e dplyr

## 2 - Pacote dplyr

- O dplyr é o pacote mais útil para realizar transformação de dados, aliando simplicidade e eficiência de uma forma elegante. Os scripts em R que fazem uso inteligente dos verbos dplyr e as facilidades do operador *pipe* (`%>%`) tendem a ficar mais legíveis e organizados sem perder velocidade de execução.
- A filosofia *tidy* é a base do [tidyverse](#). Os principais pacotes encarregados da tarefa de estruturar os dados são o dplyr e o tidyr.
- As principais funções do **dplyr** são:
  - `filter()` - filtra linhas
  - `select()` - seleciona colunas
  - `mutate()` - cria/modifica colunas
  - `arrange()` - ordena a base
  - `summarise()` - sumariza a base
- **Observação:** Esse pacote funciona nos dados em formato **tibble**. Uma tibble nada mais é do que um data.frame, mas com um método de impressão mais adequado.

# Pacote PNADcIBGE, survey e dplyr

## Aplicações no R

- ✓ Importação de dados da PNAD Contínua com a função `get_pnadc` (Dados trimestrais e anuais);
- ✓ Baixar apenas variáveis selecionadas;
- ✓ Utilizar o pacote `survey` com as seguintes funções:
  - ✓ `svytotal`
  - ✓ `svymean`
  - ✓ `svyratio`
  - ✓ `svyquantile`
  - ✓ `svyby`
  - ✓ `svyboxplot`

# Importação de Microdados da PNAD Contínua do IBGE:

**library(PNADcIBGE)**

```
get_pnadc( year = , quarter = NULL, interview =  
NULL, topic = NULL, vars = NULL, defyear =  
NULL, defperiod = NULL, labels = TRUE,  
deflator = TRUE, design = TRUE, savedir =  
tempdir() )
```

# Importação de Microdados da PNAD

## Contínua do IBGE:

- 1) **year:** O ano dos dados a serem baixados. Deve ser um número entre 2012 e o ano atual. Vetor não aceito.
- 2) **quarter:** O trimestre do ano dos dados a serem baixados. Deve ser numeração de 1 a 4. Vetor não aceito. Se NULL, o número da entrevista ou do tópico deve ser fornecido.
- 3) **interview:** O número da entrevista dos dados a serem baixados. Deve ser numeração de 1 a 5. Vetor não aceito. Usando esta opção, obterá dados anuais por entrevista. Se NULL, trimestre ou número do tópico deve ser fornecido.
- 4) **topic:** O trimestre relacionado ao tópico dos dados a serem baixados. Deve ser numeração de 1 a 4. Vetor não aceito. Usando esta opção, obterá dados anuais por tópico. Se NULL, trimestre ou número da entrevista deve ser fornecido.
- 5) **vars:** Vetor de nomes de variáveis a serem mantidos para análise. O padrão é manter todas as variáveis.



# Importação de Microdados da PNAD

## Contínua do IBGE:

- 6) **defyear:** O ano dos dados do deflator a serem baixados para microdados anuais. Deve ser um número entre 2017 e o último ano disponível. Vetor não aceito. Se NULL, o ano deflator será definido como o último ano disponível para microdados de entrevista, ou igual ao ano para microdados de tópico. Quando o trimestre for definido, este argumento será ignorado. Este argumento será usado somente se o deflator foi definido como TRUE.
- 7) **defperiod:** O período trimestral dos dados do deflator a serem baixados para microdados anuais por tópico. Deve ser numeração de 1 a 4. Vetor não aceito. Se NULL, o período deflator será definido como igual ao tópico. Quando o trimestre ou entrevista for definido, este argumento será ignorado. Este argumento será usado somente se o deflator foi definido como TRUE.
- 8) **labels:** Valor lógico. Se TRUE, as variáveis categóricas serão apresentadas como fatores com rótulos correspondentes ao dicionário da pesquisa.
- 9) **deflator:** Valor lógico. Se TRUE, as variáveis do deflator estarão disponíveis para uso nos microdados.

# Importação de Microdados da PNAD

## Contínua do IBGE:

- 10) **design**: Valor lógico. Se TRUE, retornará um objeto da classe `survey.design` ou `svyrep.design`. É altamente recomendável manter este parâmetro como TRUE para análise posterior. Se FALSE, apenas os microdados serão retornados.
- 11) **savedir**: Diretório para salvar os dados baixados. O padrão é usar um diretório temporário.

# Microdados trimestrais:

```
dadosPNADc <- get_pnadc(year = 2019, quarter = 1,  
  labels=TRUE, deflator=TRUE,  
  design=TRUE, savedir=tempdir())
```

# Microdados trimestrais:

```
dadosPNADcSelect <- get_pnadc(year = 2019, quarter = 1,  
                                vars=c("VD4001","VD4002"), labels=TRUE,  
                                deflator=TRUE, design=TRUE,  
                                savedir=tempdir())
```

# Microdados trimestrais:

# dados brutos selecionados e sem rótulo "labels = FALSE":

[illegible]

# Microdados anuais:

```
dadosPNADc_anual <- get_pnadc(year = 2019, interview = 1)
```

# Cálculo da proporção de pessoas empregadas e desempregadas:

```
library(survey)
```

```
if (!is.null(dadosPNADcSelect)) survey::svymean(x=~VD4002,  
design=pnadc.svy, na.rm=TRUE)
```

# Estimando Totais:

- A função do pacote para a estimação de totais populacionais é a **svytotal**. Sua sintaxe precisa de três parâmetros principais:
  - 1) O nome da variável que se deseja calcular o total, precedido por um ~;
  - 2) O nome do objeto do plano amostral;
  - 3) A opção `na.rm = T`, que remove as observações onde a variável é não-aplicável.
- **Observação:** o comando `keep.names = FALSE` - Evita a repetição dos labels das variáveis de forma duplicada.



# Renda Total do Brasil em 2019:

```
totalrenda <- svytotal(~VD4020, dadosPNADc, keep.names = FALSE,  
                        na.rm = TRUE)
```

**cv(totalrenda)** # Exemplo: VD4020 = 0.01752518 - Em média, os desvios atingem, em relação à média, 1,8% do valor desta.

**confint(totalrenda)** #intervalo de confiança de 95% (padrão)

**confint(totalrenda, level= .99)** #intervalo de confiança de 99%

# Variáveis Categóricas:

```
total_sexo <- svytotal(~V2007, dadosPNADc, na.rm = T)
```

```
prop_sexo <- svymean(~V2007, dadosPNADc, na.rm = T)
```

```
total_sexo_raça <- svytotal(~V2007 + V2010, dadosPNADc,  
                           keep.names = FALSE, na.rm = T)
```

```
prop_sexo_raca <- svymean(~V2007 + V2010, dadosPNADc, keep.names =  
FALSE, na.rm = T)
```

```
total_sexo_raça <- svytotal(~interaction(V2007, V2010),  
                           dadosPNADc, na.rm = T)
```

```
prop_sexo_raca <- svymean(~ interaction(V2007, V2010), dadosPNADc, na.rm =  
T)
```

# Estimando Médias:

# Rendimento mensal efetivo de todos os trabalhos

```
renda_media <- svymean(~VD4020, dadosPNADc, na.rm = T)
```

# Estimando Razões:

# Taxa de desocupação

```
txdesocup <- svyratio(~VD4002 == "Pessoas desocupadas",  
                      ~VD4001 == "Pessoas na força de trabalho",  
                      dadosPNADc, na.rm = T))
```

# Análise para o Estado do Ceará - UF=="Ceará":

# Taxa de desocupação por sexo - Ceará

```
desoc_sexo_ce <- svyby(~V2007, ~VD4002,  
                      subset(dadosPNADc, UF == "Ceará"),  
                      svymean, na.rm = T)
```

# Análise para o Estado do Ceará - UF=="Ceará":

# Rendimento médio mensal por sexo - Ceará

```
rend_med_mensal_sexo_ce <- svyby(~VD4020, ~V2007,  
                                subset(dadosPNADc, UF == "Ceará"),  
                                svymean, na.rm = T)
```

# Análise para o Estado do Ceará - UF=="Ceará":

# Distribuição de pessoas desocupadas por faixa etária

```
dadosPNADc<-update(dadosPNADc, Idade = case_when(  
  V2009 %in% 14:17 ~ "14-17",  
  V2009 %in% 18:24 ~ "18-24",  
  V2009 %in% 25:39 ~ "25-39",  
  V2009 %in% 40:59 ~ "40-59",  
  V2009 > 59 ~ "60+")  
)
```

# **Análise para o Estado do Ceará - UF=="Ceará":**

# Distribuição de pessoas desocupadas por faixa etária

```
desoc_idade_ce <- svyby(~Idade, ~VD4002,  
                        subset(dadosPNADc, UF == "Ceará"),  
                        svymean, na.rm = T)
```



# **Análise para o Estado do Ceará - UF=="Ceará":**

# Rendimento médio mensal por faixa etária - Ceará

```
rend_med_mensal_idade_ce <- svyby(~VD4020, ~Idade,  
                                   subset(dadosPNADc, UF == "Ceará"),  
                                   svymean, na.rm = T)
```