# APS 03: Amostragem e Intervalo de Confiança com PNAD Contínua

Prof. Dr. André Filipe M. Batista

15 de abril de 2025

## 1 Introdução

Esta atividade tem como objetivo aplicar conhecimentos de amostragem estatística e intervalos de confiança utilizando dados reais da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua). A análise de dados da PNAD Contínua é particularmente relevante pois proporciona uma oportunidade de trabalhar com conjuntos de dados complexos, aplicar técnicas estatísticas e de ciência de dados, e extrair insights significativos sobre a realidade socioeconômica brasileira.

## 2 A PNAD Contínua em Detalhes

### 2.1 O que é a PNAD Contínua?

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) é a principal pesquisa domiciliar do Brasil, conduzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Iniciada em 2012, ela substituiu a PNAD tradicional e a Pesquisa Mensal de Emprego (PME), consolidando a coleta de informações socioeconômicas em uma única pesquisa com cobertura nacional.

Os principais objetivos da PNAD Contínua são:

- Produzir indicadores sobre a força de trabalho (emprego, desemprego, subocupação)
- Acompanhar as flutuações trimestrais e a evolução da força de trabalho a médio e longo prazo
- Coletar dados sobre educação, migração, rendimento, condições de moradia e características demográficas
- Fornecer informações para formulação, acompanhamento e avaliação de políticas públicas
- Permitir estudos sobre desenvolvimento socioeconômico do país

### 2.2 Metodologia da Coleta de Dados

A PNAD Contínua utiliza um sofisticado plano amostral que combina várias técnicas estatísticas para garantir a representatividade da população brasileira:

#### 2.2.1 Desenho Amostral

- **Amostragem Conglomerada:** Os domicílios são selecionados em grupos (conglomerados), em vez de individualmente.
- **Amostragem Estratificada:** O país é dividido em estratos geográficos e socioeconômicos para garantir a representatividade de diferentes segmentos da população.
- **Amostragem em Múltiplos Estágios:** A seleção ocorre em etapas:
  - 1º estágio: Unidades Primárias de Amostragem (UPAs) - geralmente setores censitários
  - 2º estágio: Domicílios dentro das UPAs selecionadas
- **Esquema de Rotação de Painéis:** Cada domicílio é entrevistado por 5 trimestres consecutivos, seguindo um esquema 1-2(5), onde:

  - O domicílio é entrevistado por um mês
  - Fica fora da amostra por dois meses
  - Retorna à amostra por mais quatro vezes, seguindo o mesmo padrão
  - Total de 5 entrevistas ao longo de um período de 15 meses

#### 2.2.2 Dimensão da Amostra

- Aproximadamente 15.096 UPAs são selecionadas
- Cerca de 211.344 domicílios são entrevistados por trimestre
- A amostra cobre todos os 5.570 municípios brasileiros
- Aproximadamente 2.000 entrevistadores trabalham diariamente em campo

#### 2.2.3 Método de Coleta

- Entrevistas presenciais realizadas por técnicos do IBGE
- Utilização de dispositivos móveis de coleta (DMC) desde 2012
- Em situações especiais (como a pandemia de COVID-19), entrevistas por telefone
- Questionários estruturados com validação em tempo real
- Duração média da entrevista: 20-30 minutos

### 2.3 Sistema de Pesos Amostrais

O sistema de pesos da PNAD Contínua é particularmente sofisticado e merece atenção especial:

- **Peso Básico:** Calculado como o inverso da probabilidade de seleção do domicílio.
- **Ajuste de Não-Resposta:** Os pesos são ajustados para compensar domicílios selecionados mas não entrevistados (recusas, ausências, etc.).
- **Calibração:** Os pesos são calibrados para que as estimativas da pesquisa coincidam com os totais populacionais conhecidos (projeções demográficas) por:
  - Sexo
  - Grupos etários
  - Unidades da Federação
- **Pesos Longitudinais:** Para análises que acompanham os mesmos domicílios ao longo do tempo, são calculados pesos específicos para controlar o atrito (perda de unidades amostrais ao longo do tempo).

Na prática, o peso amostral é representado pela variável **V1028** nos microdados. Este é o peso que deve ser utilizado em todas as análises para obter estimativas que representem corretamente a população brasileira.

## 3 Atividade Prática: Análise da PNAD Contínua

### 3.1 Objetivos de Aprendizagem

Ao completar esta atividade, os grupos serão capazes de:

- Manipular e analisar microdados de pesquisas amostrais complexas
- Aplicar conceitos de amostragem estratificada e conglomerada
- Utilizar pesos amostrais para obter estimativas não enviesadas
- Calcular intervalos de confiança utilizando a técnica de bootstrap
- Reproduzir estatísticas oficiais a partir de microdados
- Interpretar resultados estatísticos no contexto socioeconômico brasileiro

## 3.2 Acesso aos Dados

Os microdados da PNAD Contínua estão disponíveis no site do IBGE: `https://www.ibge.gov.br/estatisticas/sociais/trabalho/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html`

Para esta atividade, utilize os dados do último trimestre disponível. Você precisará baixar:

- Arquivo de microdados (.csv ou .txt)
- Dicionário de variáveis (.xls ou .pdf)
- Nota técnica sobre o cálculo dos pesos (disponível na documentação)

**Dica:** A biblioteca Python `Pnadium` facilita a manipulação de microdados da PNAD: `https://www.linkedin.com/pulse/pnadium-um-pacote-para-manusear-microdados-da-pnad-contínua-ximenez-lku2f/`

**Atenção:** Dependendo do trimestre de dados que você captura, a variável indicada pode não existir ou estar codificada sob um novo código. É responsabilidade do grupo verificar o dicionário de dados e achar a variável mais indicada. Na eventual inexistência da variável na base de dados, o grupo deve propor uma nova análise e justificar.

**RECOMENDA-SE FORTEMENTE QUE OS GRUPOS TRABALHEM COM OS DADOS DO 4o. TRIMESTRE DE 2024, PORÉM FICA LIVRE AO GRUPO TRABALHAR COM OUTRO PERÍODO**.

## 3.3 Guia Passo a Passo

### 3.3.1 1. Preparação do Ambiente

- Instale as bibliotecas necessárias: pandas, numpy, matplotlib, seaborn e statsmodels
- Configure seu ambiente Python para trabalhar com grandes conjuntos de dados
- Familiarize-se com a documentação da PNAD Contínua

### 3.3.2 2. Carregamento e Exploração dos Dados

- Carregue os microdados usando pandas ou pnadium
- Explore a estrutura dos dados (tipos de variáveis, valores ausentes, distribuições)
- Identifique e decodifique as variáveis relevantes para seu grupo

### 3.3.3 3. Filtragem e Transformação

- Filtre a população de interesse de acordo com a definição do IBGE
- Crie as variáveis derivadas necessárias para sua análise
- Trate adequadamente os valores ausentes e extremos

### 3.3.4 4. Cálculo das Estatísticas Ponderadas

- Use a variável de peso amostral (V1028) em todos os cálculos
- Para médias ponderadas, use `np.average()` com o parâmetro `weights`
- Para proporções, use a função apropriada do statsmodels ou calcule manualmente
- Calcule as estatísticas para o Brasil e para os recortes específicos do seu grupo

### 3.3.5 5. Cálculo de Intervalos de Confiança por Bootstrap

- Implemente a técnica de bootstrap para estimar intervalos de confiança
- Execute o bootstrap considerando o desenho amostral complexo
- Calcule os intervalos de confiança de 95% para todas as estatísticas

#### 3.3.6 6. Validação com Estatísticas Oficiais

- Pesquise as estatísticas oficiais divulgadas pelo IBGE para o período analisado
- Compare suas estimativas com os valores oficiais
- Calcule e interprete as diferenças encontradas

#### 3.3.7 7. Visualização e Apresentação dos Resultados

- Crie gráficos e tabelas que apresentem suas estimativas e intervalos de confiança
- Compare visualmente os diferentes recortes populacionais
- Prepare uma apresentação ou relatório com os resultados e interpretações

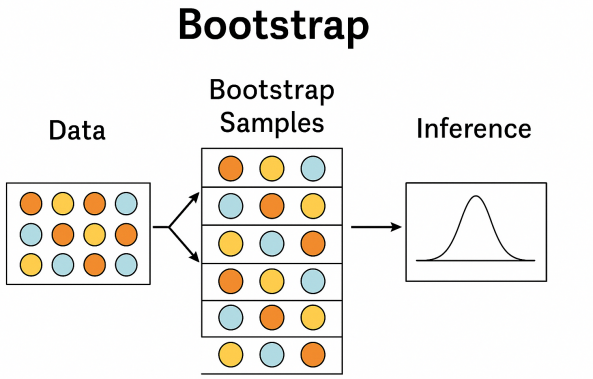
# 4 Entendendo o Bootstrap para Intervalos de Confiança

Uma parte importante desta atividade é o cálculo de intervalos de confiança utilizando a técnica de bootstrap. Como a PNAD Contínua tem um desenho amostral complexo, o cálculo convencional de intervalos de confiança pode ser impreciso. O bootstrap permite estimar a variabilidade das estatísticas sem fazer suposições sobre a distribuição dos dados ou ignorar o desenho amostral.

## 4.1 O que é o Bootstrap?

O bootstrap é uma técnica estatística de reamostragem introduzida por Bradley Efron em 1979. A ideia central é simples mas poderosa: em vez de fazer suposições teóricas sobre a distribuição da estatística de interesse, utilizamos os próprios dados para simular essa distribuição.

O bootstrap funciona criando múltiplas "amostras bootstrap" a partir da amostra original. Cada amostra bootstrap é gerada através de amostragem com reposição da amostra original, mantendo o mesmo tamanho. Ao calcular a estatística de interesse em cada uma dessas amostras bootstrap, obtemos uma distribuição empírica dessa estatística, que pode ser usada para calcular intervalos de confiança.




## 4.2 Por que usar Bootstrap com a PNAD Contínua?

A PNAD Contínua utiliza um desenho amostral complexo, que inclui estratificação, conglomeração e pesos amostrais. Esse desenho viola as premissas de amostragem aleatória simples assumidas por muitos métodos estatísticos convencionais. Consequências de ignorar o desenho amostral incluem:

- Subestimação dos erros padrão
- Intervalos de confiança muito estreitos
- Testes estatísticos com taxas de erro Tipo I inflacionadas
- Inferências incorretas sobre a população

O bootstrap oferece uma solução robusta para esse problema, permitindo incorporar o desenho amostral complexo no processo de reamostragem.

# 5 Bootstrap para Intervalos de Confiança: Implementação Prática

O bootstrap é uma técnica estatística que permite estimar a variabilidade de estatísticas sem fazer suposições sobre sua distribuição. Essa técnica é especialmente valiosa quando trabalhamos com pesquisas amostrais complexas como a PNAD Contínua.

## 5.1 Função Bootstrap pronta para uso

Abaixo está uma função pronta para uso que implementa o bootstrap para calcular intervalos de confiança considerando os pesos amostrais da PNAD Contínua.

```
1
2  def bootstrap_ci(df, col, weight_col, n_bootstrap=1000, alpha=0.05):
3      """
4      Calcula estatística (média ou proporção ponderada) e intervalo de confiança usando
           bootstrap.
5
6      Parâmetros:
7      -----------
8      df : DataFrame
9          DataFrame contendo os dados
10     col : str
11         Nome da coluna com a variável numérica (pode ser 0/1 ou contínua)
12     weight_col : str
13         Nome da coluna contendo os pesos amostrais
14     n_bootstrap : int, opcional
15         Número de amostras bootstrap (padrão: 1000)
16     alpha : float, opcional
17         Nível de significância (padrão: 0.05 para IC de 95%)
18
19     Retorna:
20     --------
21     float, tuple
22         Estatística ponderada original e tupla com limites inferior e superior do IC
23     """
24     # Estatística original (proporção ou média ponderada)
25     stat_original = np.average(df[col], weights=df[weight_col])
26
27     # Estimativas bootstrap
28     estimates = []
29
30     for _ in range(n_bootstrap):
31         sample = df.sample(frac=1, replace=True)
32         stat = np.average(sample[col], weights=sample[weight_col])
33         estimates.append(stat)
34
35     # Intervalo de confiança
36     lower = np.percentile(estimates, 100 * alpha / 2)
37     upper = np.percentile(estimates, 100 * (1 - alpha / 2))
38
39     return stat_original, (lower, upper)
```

Listing 1: Funções Bootstrap para Proporções e Médias com Pesos Amostrais

## 5.2 Interpretando os Resultados do Bootstrap

Após executar o procedimento bootstrap, você obterá:

- **Estimativa pontual:** O valor da estatística calculada na amostra original
- **Distribuição bootstrap:** A distribuição da estatística nas amostras bootstrap
- **Intervalo de confiança:** Os limites que contêm a estatística populacional com a probabilidade especificada (geralmente 95%)

A distribuição bootstrap pode ser visualizada em um histograma, mostrando a variabilidade da estatística. Quanto mais estreita essa distribuição, maior a precisão da estimativa.

Para interpretar o intervalo de confiança, lembre-se que um intervalo de confiança de 95% significa que, se repetíssemos o processo de amostragem e cálculo do intervalo muitas vezes, 95% desses intervalos conteriam o verdadeiro valor populacional.

# 6 Instruções Detalhadas

## Distribuição de UF's e Regiões por Grupo

Cada grupo utilizará sempre as mesmas Unidades da Federação (UFs) e Regiões para todas as análises solicitadas.

| Grupo | UFs (siglas) | Regiões |
|---|---|---|
| 1 | SP, MG, RJ | Sudeste, Sul |
| 2 | BA, PE, CE | Nordeste, Norte |
| 3 | RS, SC, PR | Sul, Sudeste |
| 4 | PA, AM, RO | Norte, Centro-Oeste |
| 5 | DF, GO, MT | Centro-Oeste, Nordeste |
| 6 | MA, PI, PB | Nordeste, Norte |
| 7 | ES, MS, TO | Sudeste, Centro-Oeste |
| 8 | AL, SE, RN | Nordeste, Sul |

Tabela 1: Distribuição das UF's e Regiões por grupo

## 6.1 Análise #1: Rendimento Médio Real e Taxa de Desemprego

### 6.1.1 Estatística Numérica: Rendimento Médio Real (VD4020)

Para calcular o rendimento médio real dos trabalhadores, siga estas etapas:

1. **Identificação das variáveis relevantes:**
   - VD4020: Rendimento mensal habitual do trabalho principal
   - VD4002: Condição em relação à força de trabalho (1 = ocupadas, 2 = desocupadas)
   - V2007: Sexo (1 = homem, 2 = mulher)
   - V1028: Peso amostral
2. **Filtragem dos dados:**
   - Filtre apenas pessoas ocupadas (VD4002 = 1)
   - Remova registros com rendimento inválido ou não informado (VD4020 $\leq 0$)
   - Crie uma cópia do DataFrame filtrado para evitar problemas de referência
3. **Cálculo do rendimento médio para todo o Brasil:**
   - Utilize a média ponderada pelo peso amostral (V1028)
   - Aplique o método bootstrap para calcular o intervalo de confiança de 95%
4. **Cálculo do rendimento médio por sexo:**
   - Filtre os dados separadamente para homens (V2007 = 1) e mulheres (V2007 = 2)
   - Para cada grupo, calcule a média ponderada pelo peso amostral (V1028)
   - Aplique o método bootstrap para calcular o intervalo de confiança de 95%
5. **Visualização dos resultados:**
   - Crie um gráfico de barras comparando o rendimento médio para as UFs do seu grupo e por sexo
   - Inclua barras de erro representando os intervalos de confiança

- Adicione títulos, rótulos e legenda informativos

6. **Comparação com dados oficiais:**
   - Consulte o SIDRA (tabela 5442) ou os relatórios trimestrais do IBGE
   - Compare suas estimativas com as estatísticas oficiais
   - Calcule o percentual de diferença e discuta possíveis causas

#### 6.1.2 Estatística de Proporção: Taxa de Desemprego

Para calcular a taxa de desemprego, siga estas etapas:

1. **Identificação das variáveis relevantes:**
   - VD4002: Condição em relação à força de trabalho (1 = ocupadas, 2 = desocupadas)
   - V2009: Idade do morador
   - V2007: Sexo (1 = homem, 2 = mulher)
   - V1028: Peso amostral
2. **Filtragem dos dados:**
   - Filtre apenas pessoas em idade de trabalhar (V2009 $\geq$ 14)
   - Crie uma variável binária 'forca_trabalho' que identifica pessoas ocupadas ou desocupadas (VD4002 = 1 ou 2)
   - Crie uma variável binária 'desocupado' que identifica apenas pessoas desocupadas (VD4002 = 2)
   - Filtre apenas pessoas na força de trabalho (forca_trabalho = 1)
3. **Cálculo da taxa de desemprego para as suas regiões:**
   - Calcule a proporção ponderada de pessoas desocupadas entre as pessoas na força de trabalho
   - Use o peso amostral (V1028) em todos os cálculos
   - Aplique o método bootstrap para calcular o intervalo de confiança de 95%
   - Multiplique por 100 para expressar como percentagem
4. **Cálculo da taxa de desemprego por sexo:**
   - Filtre os dados separadamente para homens (V2007 = 1) e mulheres (V2007 = 2)
   - Para cada grupo, calcule a proporção ponderada de desocupados
   - Aplique o método bootstrap para calcular os intervalos de confiança de 95%
5. **Visualização dos resultados:**
   - Crie um gráfico de barras comparando a taxa de desemprego para o Brasil e por sexo
   - Inclua barras de erro representando os intervalos de confiança
   - Adicione títulos, rótulos e legenda informativos
6. **Comparação com dados oficiais:**
   - Consulte o SIDRA ou os relatórios trimestrais do IBGE
   - Compare suas estimativas com as estatísticas oficiais
   - Calcule a diferença em pontos percentuais e discuta possíveis causas

## 6.2 Análise #2: Horas Médias Trabalhadas e Taxa de Informalidade

### 6.2.1 Estatística Numérica: Horas Médias Trabalhadas por Semana (VD4031)

**Atenção:** Essa análise faremos para todas as regiões, mas no relatório o seu grupo deve focar em interpretar as regiões que foram atribuídas ao grupo.

Para calcular as horas médias trabalhadas por semana, siga estas etapas:

1. **Identificação das variáveis relevantes:**
   - VD4031: Horas habitualmente trabalhadas por semana em todos os trabalhos
   - VD4002: Condição em relação à força de trabalho (1 = ocupadas)
   - UF: Unidade da Federação (códigos de 11 a 53)
   - V1028: Peso amostral
2. **Filtragem dos dados:**
   - Filtre apenas pessoas ocupadas (VD4002 = 1)
   - Remova registros com horas inválidas (VD4031 = 0 ou valores muito altos, acima de 98)
   - Crie uma cópia do DataFrame filtrado para evitar problemas de referência
3. **Criação da variável de região:**
   - Crie uma variável 'regiao' a partir da UF seguindo a divisão oficial do IBGE:
     - Norte: UF $\in$ {11, 12, 13, 14, 15, 16, 17}
     - Nordeste: UF $\in$ {21, 22, 23, 24, 25, 26, 27, 28, 29}
     - Sudeste: UF $\in$ {31, 32, 33, 35}
     - Sul: UF $\in$ {41, 42, 43}
     - Centro-Oeste: UF $\in$ {50, 51, 52, 53}
4. **Cálculo das horas médias para todo o Brasil:**
   - Utilize a média ponderada pelo peso amostral (V1028)
   - Aplique o método bootstrap para calcular o intervalo de confiança de 95%
5. **Cálculo das horas médias por região:**
   - Agrupe os dados por região
   - Para cada região, calcule a média ponderada pelo peso amostral (V1028)
   - Aplique o método bootstrap para calcular os intervalos de confiança de 95%
6. **Visualização dos resultados:**
   - Crie um gráfico de barras comparando as horas médias por região
   - Inclua uma linha horizontal indicando a média nacional
   - Adicione barras de erro representando os intervalos de confiança
7. **Comparação com dados oficiais:**
   - Consulte o SIDRA (tabela 5453) ou relatórios do IBGE
   - Compare suas estimativas com as estatísticas oficiais
   - Calcule as diferenças e discuta possíveis causas

### 6.2.2 Estatística de Proporção: Taxa de Informalidade

Para calcular a taxa de informalidade, siga estas etapas:

1. **Identificação das variáveis relevantes:**
   - VD4002: Condição em relação à força de trabalho (1 = ocupadas)
   - VD4009: Posição na ocupação no trabalho principal
   - V4019: CNPJ (1 = tem, 2 = não tem)
   - UF: Unidade da Federação
   - V1028: Peso amostral
2. **Filtragem dos dados:**
   - Filtre apenas pessoas ocupadas (VD4002 = 1)
3. **Criação da variável de informalidade:**
   - Trabalhadores formais são aqueles que têm marcadores de formalidade. São normalmente empregados três marcadores de formalidade, sendo estes: (i) carteira assinada para empregados do setor público e privado; (ii) Militares e estatutários para empregados do setor público; e (iii) Empresa com CNPJ, para empregadores e autônomos.
   - Trabalhadores que não se enquadram em nenhuma dessas três categorias são considerados informais.
   - Crie uma variável binária 'informal' com valor 1 para trabalhadores informais e 0 para formais
   - Considere formais:
     - Ocupados (VD4001 = 1 e VD4002 = 1) e empregados com carteira de trabalho assinada (VD4009 = 1, 3 ou 5) ou militares ou estatutários (VD4009 = 7)
     - Ocupados (VD4001 = 1 e VD4002 = 1) e empregador ou conta própria (VD4009 = 8 e 9) e o negócio possui CNPJ (V4019 = 1).
4. **Criação da variável de região:** (igual ao item anterior)
5. **Cálculo da taxa de informalidade para todo o Brasil:**
   - Calcule a proporção ponderada de trabalhadores informais entre os ocupados
   - Use o peso amostral (V1028) em todos os cálculos
   - Aplique o método bootstrap para calcular o intervalo de confiança de 95%
   - Multiplique por 100 para expressar como percentagem
6. **Cálculo da taxa de informalidade por região:**
   - Agrupe os dados por região
   - Para cada região, calcule a proporção ponderada de trabalhadores informais
   - Aplique o método bootstrap para calcular os intervalos de confiança de 95%
7. **Visualização dos resultados:**
   - Crie um gráfico de barras comparando a taxa de informalidade por região
   - Inclua uma linha horizontal indicando a média nacional
   - Adicione barras de erro representando os intervalos de confiança
   - Opcionalmente, crie um mapa coroplético do Brasil (neste caso, faça do Brasil todo, não apenas da região do seu grupo) destacando as regiões por taxa de informalidade

## 6.3 Análise #3: Anos Médios de Estudo e Proporção de Jovens Nem-Nem

### 6.3.1 Estatística Numérica: Anos Médios de Estudo da População Ocupada

Para calcular os anos médios de estudo, siga estas etapas:

1. **Identificação das variáveis relevantes:**
   - VD3004: Nível de instrução mais elevado alcançado
   - VD3005: Anos de estudo padronizado para o ensino fundamental
   - VD4002: Condição em relação à força de trabalho (1 = ocupadas)
   - V2007: Sexo (1 = homem, 2 = mulher)
   - V1028: Peso amostral
2. **Filtragem dos dados:**
   - Filtre apenas pessoas ocupadas (VD4002 = 1)
3. **Cálculo dos anos médios de estudo para as suas regiões:**
   - Utilize a média ponderada pelo peso amostral (V1028)
   - Aplique o método bootstrap para calcular o intervalo de confiança de 95%
4. **Cálculo dos anos médios de estudo por sexo:**
   - Filtre os dados separadamente para homens (V2007 = 1) e mulheres (V2007 = 2)
   - Para cada grupo, calcule a média ponderada pelo peso amostral
   - Aplique o método bootstrap para calcular os intervalos de confiança de 95%
5. **Visualização dos resultados:**
   - Crie um gráfico de barras comparando os anos médios de estudo por sexo
   - Inclua uma barra para a média nacional
   - Adicione barras de erro representando os intervalos de confiança

### 6.3.2 Estatística de Proporção: Jovens Nem-Nem (18-29 anos)

Para calcular a proporção de jovens nem-nem, siga estas etapas:

1. **Identificação das variáveis relevantes:**
   - V2009: Idade do morador
   - VD4002: Condição em relação à força de trabalho (1 = ocupadas, 2 = desocupadas)
   - V3002: Frequenta escola (1 = sim, 2 = não)
   - V2007: Sexo (1 = homem, 2 = mulher)
   - V1028: Peso amostral
2. **Filtragem dos dados:**
   - Filtre apenas jovens entre 18 e 29 anos (18  V2009  29)
3. **Criação da variável nem-nem:**
   - Crie uma variável 'estuda' onde estuda = 1 se V3002 = 1, caso contrário 0
   - Crie uma variável 'trabalha' onde trabalha = 1 se VD4002 = 1 (ocupado), caso contrário 0
   - Crie uma variável 'nem_nem' = 1 se estuda = 0 e trabalha = 0, caso contrário 0
4. **Cálculo da proporção de nem-nem para as suas regiões:**
   - Calcule a proporção ponderada de jovens nem-nem entre os jovens de 18-29 anos
   - Use o peso amostral (V1028) em todos os cálculos

- Aplique o método bootstrap para calcular o intervalo de confiança de 95%
- Multiplique por 100 para expressar como percentagem

5. **Cálculo da proporção de nem-nem por sexo:**
   - Filtre os dados separadamente para homens (V2007 = 1) e mulheres (V2007 = 2)
   - Para cada grupo, calcule a proporção ponderada de jovens nem-nem
   - Aplique o método bootstrap para calcular os intervalos de confiança de 95%
6. **Visualização dos resultados:**
   - Crie um gráfico de barras comparando a proporção de nem-nem por sexo
   - Inclua uma barra para a média nacional
   - Adicione barras de erro representando os intervalos de confiança
7. **Análise adicional:**
   - Calcule a proporção de jovens que apenas estudam, apenas trabalham, estudam e trabalham, e nem-nem
   - Crie um gráfico de pizza ou barras empilhadas para visualizar essa distribuição

## 6.4 Análise #4: Estatística de Proporção: Pessoas Abaixo da Linha de Pobreza

Para calcular a proporção de pessoas abaixo da linha de pobreza, siga estas etapas:

1. **Identificação das variáveis relevantes:**
   - VD4019: Rendimento domiciliar per capita
   - UF: Unidade da Federação
   - V1028: Peso amostral da pessoa
   - V4019: Rendimento mensal efetivo de todos os trabalho

   Para calcular o rendimento médio per capita poderado, você deve:
   - Criar uma variável 'id_domicilio', formada pela concatenação das variáveis 'UPA', 'V1008' e 'V1014'
   - Agrupar por identificador do domicílio a variável VD4019, somando-a e obtendo o valor de renda total
   - Obter o numero de moradores por domicílio
   - Calcular a renda per capita ponderada pelo peso amostral
2. **Definição da linha de pobreza:**
   - Identifique o valor do salário mínimo vigente no período da PNAD
   - Defina a linha de pobreza como meio salário mínimo per capita
   - Crie uma variável binária 'pobre' = 1 se Renda Media per capita < (salário mínimo / 2), caso contrário 0
3. **Cálculo da proporção para suas regiões:**
   - Calcule a proporção ponderada de pessoas abaixo da linha de pobreza
   - Use o peso específico em todos os cálculos
   - Aplique o método bootstrap para calcular o intervalo de confiança de 95%
   - Multiplique por 100 para expressar como percentagem
4. **Cálculo da proporção por UF:**

- Agrupe os dados por UF
- Para cada UF, calcule a proporção ponderada de pessoas abaixo da linha de pobreza
- Aplique o método bootstrap para calcular os intervalos de confiança de 95%

5. **Visualização dos resultados:**
   - Crie um gráfico de barras ordenado (do maior para o menor) comparando a taxa de pobreza por UF
   - Inclua uma linha horizontal indicando a média nacional
   - Adicione barras de erro representando os intervalos de confiança
   - Desafio: Crie um mapa coroplético do Brasil destacando as UFs por taxa de pobreza

# Rubrica de Avaliação e Entrega do Relatório

## Instruções Gerais para o Relatório

Cada grupo deverá entregar um relatório técnico contendo:

- **Introdução:** contextualização do problema, justificativa para as variáveis escolhidas e recortes regionais adotados.
- **Metodologia:** descrição clara do processo de tratamento dos dados, filtragem, criação de variáveis e uso dos pesos amostrais.
- **Resultados:** apresentação dos resultados com tabelas e gráficos, incluindo as estimativas pontuais e intervalos de confiança.
- **Validação:** comparação com estatísticas oficiais do IBGE e discussão sobre as diferenças encontradas.
- **Conclusão:** interpretação crítica dos resultados e limitações da abordagem.
- **Código-fonte:** anexo ou link para repositório com scripts completos e reprodutíveis.

## Critérios de Avaliação

| **Critério** | **Peso** | **Nota (0-10)** |
|---|---|---|
| Clareza e estrutura do relatório | 10% | |
| Correção na manipulação dos dados e uso do peso amostral | 20% | |
| Implementação correta do método bootstrap | 20% | |
| Apresentação visual dos resultados (gráficos e tabelas) | 15% | |
| Validação com dados oficiais do IBGE e discussão crítica | 15% | |
| Código limpo, comentado e reprodutível | 10% | |
| Entrega pontual e organização do material | 10% | |
| **Total** | **100%** | |

Tabela 2: Rubrica de Avaliação da APS 03