

O uso do método bootstrap para cálculo de estimativas de variância na PNAD Contínua

Marcus Vinicius M. Fernandes

PNAD Contínua

- Principal pesquisa do Sistema Integrado de Pesquisas Domiciliares do IBGE;
- Seu principal objetivo é produzir informações básicas para o estudo socioeconômico do país e permitir a investigação contínua de indicadores sobre mercado de trabalho;
- Realizada pelo IBGE desde o início de 2012.

O que é divulgado com a PNAD Contínua?

- Divulgações mensais (trimestre móvel) de alguns indicadores de mercado de trabalho e rendimento – somente para Brasil ;
- Divulgações trimestrais sobre mercado de trabalho e rendimento para Brasil, Grandes Regiões, Unidades da Federação, Regiões Metropolitanas e Capitais;
- Divulgações anuais de alguns temas como:
 - Trabalho Infantil;
 - Educação;
 - Outras formas de trabalho;
 - Rendimentos;
 - Acesso à Internet.

Plano Amostral da PNAD Contínua

- Amostra estratificada e conglomerada com seleção de UPAS (Unidades Primárias de Amostragem), com probabilidade proporcional ao tamanho – número de domicílios.

- Estratificação:
 - Estratificação por divisão administrativa;
 - Estratificação geográfica especial;
 - Estratificação por situação dos domicílios;
 - Estratificação Estatística.

Plano Amostral da PNAD Contínua

- Plano amostral conglomerado em dois estágios:
 - **1º Estágio:** são selecionadas UPAS com probabilidade proporcional ao número de domicílios dentro de cada estrato;
 - **2º Estágio:** número fixo de domicílios particulares permanentes ocupados, através de amostra aleatória simples do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE) atualizado.

- Foi fixado em 14 o número de domicílios selecionados em cada UPA;

- Cerca de 15.100 UPAS / 3.465 municípios / 211.000 domicílios – por trimestre.

Esquema de rotação da PNAD Contínua

- Utiliza amostra com painel rotacional;
- Esquema de rotação de domicílios: 1– 2(5);
 - 80% de sobreposição de amostra no trimestre;
 - 20% de sobreposição de amostra no ano;
- Esquema de rotação de UPAS: 2,5% a cada trimestre.
 - Após 10 anos renovação total de UPAS.

Estimação na PNAD Contínua

- Fatores de expansão associados a UPAs, domicílios e seus moradores:
 - **Peso básico** : proveniente do desenho da amostra;
 - **Peso com ajuste por não resposta** : trata a perda como sendo aleatória e uniforme dentro das Unidades Primárias de Amostragem (UPA) da pesquisa ;
 - **Peso ajustado por pós-estratificação**: coerência com as estimativas populacionais – 77 pós-estratos geográficos.

Cálculo de estimativa de variância na PNAD Contínua

- O IBGE publica estimativas de Coeficiente de Variação (CV) para todos os indicadores divulgados e para o cálculo das estimativas de variância da pesquisa, aplica-se o método do conglomerado primário com linearização de Taylor;
- **Recentemente:** implementação de novo método de calibração. Método escolhido foi o raking, utilizando totais populacionais por regiões geográficas, sexo e classes de idade (total para o país);
- Escolha do método para o cálculo da variância: linearização de Taylor x métodos de replicação; escolha de software para implementação (SAS-SUDAAN x R);
- Opção pelo método de replicação **bootstrap**.

O método de bootstrap

- Rao and Wu (1988) e Rao, Wu e Yue (1992) – método de replicação de pesos amostrais para cálculo da variância;

- O método:

Seja o interesse estimar a variância da estimador $\hat{\theta}$ de um parâmetro de interesse θ :

- a) Seleciona-se uma amostra aleatória simples, com reposição, de $m_h = n_h - 1$ UPAs das n_h UPAS existentes no estrato H;
- b) Repita o passo a) por um número grande de vezes, digamos B, e seja a quantidade $m_{hi}(b)$, o número de vezes que a junção estrato h, UPA i, aparecem na b-ésima amostra bootstrap;

O método de bootstrap

- c) O peso bootstrap para uma unidade k é calculado, então, da seguinte maneira:

$$w_{hik}(b) = \frac{w_{hik}n_h m_{hi}(b)}{n_h - 1}$$

Onde w_{hik} é o peso da unidade k , da UPA i , do estrato h .

- d) Recalcula-se $\hat{\theta}$ usando os pesos bootstrap, $\hat{\theta}_{(b)}$, para $b = 1, \dots, B$;
- e) A variância bootstrap pode ser definida por:

$$v_B = \frac{\sum_{b=1}^B (\hat{\theta}_{(b)} - \bar{\theta}_{(\cdot)})^2}{B}$$

Onde:
$$\bar{\theta}_{(\cdot)} = \frac{\sum_b \hat{\theta}_{(b)}}{B}$$

O método de bootstrap

- A estimativa $\bar{\theta}_{(.)}$ pode ser substituída pela estimativa pontual;
- A quantidade B de pesos replicados pode ser escolhida de acordo com a convergência e a necessidade de cada pesquisa;
- Quando ajustes de não-resposta e calibração são feitos nos pesos originais da pesquisa, estes devem ser realizados também em cada uma das amostras bootstrap;

Resultados

Tabela: Coeficientes de variação de alguns indicadores da PNAD Contínua – linearização de Taylor x bootstrap - usando pós-estratificação - 1º trimestre de 2014.

	Pessoas de 14 anos ou mais (%)		Pessoas na força de trabalho (%)	
	Linearização	Bootstrap	Linearização	Bootstrap
Brasil	0,114	0,115	0,222	0,218
Grandes Regiões				
Regiao Norte	0,318	0,315	0,578	0,580
Regiao Nordeste	0,191	0,187	0,409	0,427
Regiao Sudeste	0,209	0,208	0,400	0,397
Regiao Sul	0,217	0,207	0,424	0,401
Regiao Centro Oeste	0,297	0,295	0,521	0,502

Tabela: Coeficientes de variação de alguns indicadores da PNAD Contínua – linearização de Taylor x bootstrap - usando pós-estratificação - 1º trimestre de 2014.

	Pessoas ocupadas (%)		Taxa de desocupação	
	Linearização	Bootstrap	Linearização	Bootstrap
Brasil	0,241	0,242	1,234	1,179
Grandes Regiões				
Regiao Norte	0,618	0,622	2,979	2,911
Regiao Nordeste	0,451	0,483	1,745	1,754
Regiao Sudeste	0,436	0,435	2,337	2,225
Regiao Sul	0,443	0,421	2,855	2,965
Regiao Centro Oeste	0,562	0,539	3,237	3,166

Tabela: Coeficientes de variação de alguns indicadores da PNAD Contínua – linearização de Taylor x bootstrap - usando calibração (raking) por sexo e idade - 1º trimestre de 2014.

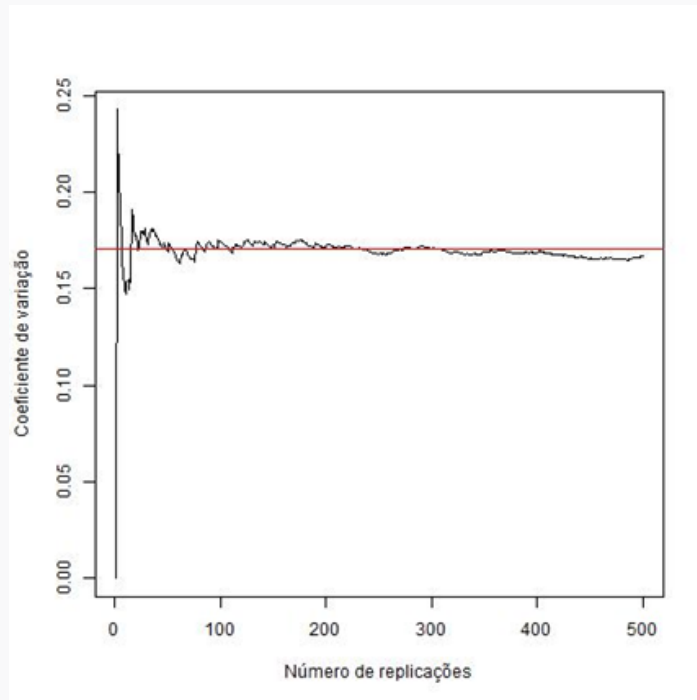
	Pessoas de 14 anos ou mais (%)		Pessoas na força de trabalho (%)	
	Survey	Bootstrap	Survey	Bootstrap
Brasil	0,000	0,000	0,171	0,167
Grandes Regiões				
Regiao Norte	0,345	0,347	0,583	0,602
Regiao Nordeste	0,195	0,183	0,393	0,379
Regiao Sudeste	0,156	0,154	0,354	0,351
Regiao Sul	0,240	0,233	0,415	0,411
Regiao Centro Oeste	0,332	0,323	0,524	0,485

Tabela: Coeficientes de variação de alguns indicadores da PNAD Contínua – linearização de Taylor x bootstrap - usando calibração (raking) por sexo e idade - 1º trimestre de 2014.

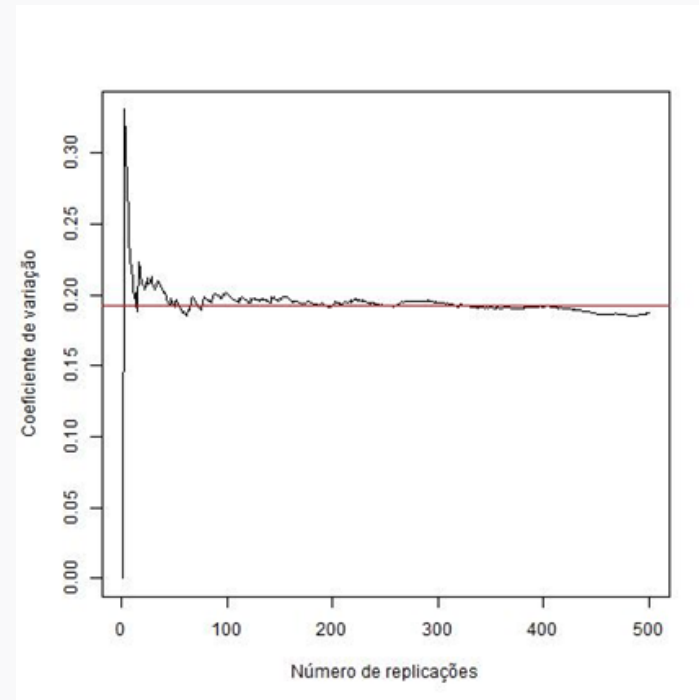
	Pessoas ocupadas (%)		Taxa de desocupação	
	Survey	Bootstrap	Survey	Bootstrap
Brasil	0,193	0,187	1,210	1,220
Grandes Regiões				
Regiao Norte	0,620	0,641	2,998	2,951
Regiao Nordeste	0,435	0,414	1,755	1,723
Regiao Sudeste	0,396	0,384	2,307	2,230
Regiao Sul	0,436	0,436	2,892	2,986
Regiao Centro Oeste	0,567	0,543	3,262	3,290

Figura: Gráficos de convergência das amostras bootstrap – Coeficientes de variação (pesos com calibração), para alguns indicadores selecionados – PNAD Contínua - 1º trimestre de 2014:

(a) Pessoas na força de trabalho



(b) Pessoas ocupadas



Algumas definições para a PNAD Contínua

- Calibração raking utilizando o pacote survey do R – aplicando algumas restrições;
- Geração das amostras bootstrap (200 replicações) – usando o método (n-1) também utilizando o pacote survey do R;
- Estratos com apenas uma UPA são agregados;
- Calibração das 200 amostras bootstrap, utilizando o mesmo modelo e restrições;

Algumas definições para a PNAD Contínua

- Serão disponibilizados no microdado de uso público: o peso final calibrado e as 200 replicações – simplicidade para usuários calcularem erros amostrais;
- Previsão de divulgação ainda este ano (setembro);
- Estudos relativos à estimativa de variância de diferenças entre períodos (amostras bootstrap coordenadas).

Obrigado

marcus.fernandes@ibge.gov.br