UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
GRADUAÇÃO EM ESTATÍSTICA

MARIA DO CARMO TEIXEIRA TOCANTINS ALVIM

COMBINAÇÃO DE DADOS AMOSTRAIS DE DUAS EDIÇÕES DA PNAD CONTÍNUA

JUIZ DE FORA

2023

MARIA DO CARMO TEIXEIRA TOCANTINS ALVIM

COMBINAÇÃO DE DADOS AMOSTRAIS DE DUAS EDIÇÕES DA PNAD CONTÍNUA

Monografia apresentada pelo(a) acadêmico(a) Maria do Carmo Teixeira Tocantins Alvim ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Marcel de Toledo Vieira, Ph.D.

Juiz de Fora

2023

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Alvim, Maria do Carmo Teixeira Tocantins.
COMBINAÇÃO DE DADOS AMOSTRAIS DE DUAS EDIÇÕES DA PNAD CONTÍNUA / Maria do Carmo Teixeira Tocantins Alvim. -- 2023.
46 f.

Orientador: Marcel de Toledo Vieira
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2023.

1. Combinação de Amostras. 2. Abordagens combinada e separada. 3. Amostragem complexa. 4. Pnad Contínua. I. Vieira, Marcel de Toledo, orient. II. Título.

MARIA DO CARMO TEIXEIRA TOCANTINS ALVIM

COMBINAÇÃO DE DADOS AMOSTRAIS DE DUAS EDIÇÕES DA PNAD CONTÍNUA

Monografia apresentada pelo(a) acadêmico(a) Maria do Carmo Teixeira Tocantins Alvim ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para obtenção do título de Bacharel em Estatística.

Aprovada em 14 de julho de 2023.

BANCA EXAMINADORA

Prof. Marcel de Toledo Vieira, Ph.D. - Orientador
Universidade Federal de Juiz de Fora

Prof.ª Dr.ª Ângela Mello Coelho
Universidade Federal de Juiz de Fora

Prof. Dr. Augusto Carvalho Souza
Universidade Federal de Juiz de Fora

Agradecimentos

Em primeiro lugar, agradeço a Deus por ter me dado o direito à vida e, com ela, dons especiais que me fizeram e fazem ser uma pessoa melhor. Aos meus pais (*in memoriam*), pela criação e exemplo de vida.

Agradeço aos professores, por terem compartilhado seus conhecimentos, ensinando que não existe nenhuma fronteira para o saber. E, em especial, ao Professor Marcel, pela sua dedicação, paciência e confiança em minha pessoa para a realização deste trabalho.

Agradeço a minhas filhas Mariana e Carolina, pelo incentivo dado objetivando vencer os desafios imposto pelo mundo e pela compreensão de muitas vezes não terem minha presença em alguns momentos de suas vidas.

Agradeço a meu esposo, Paulo Roberto, pela ajuda com as tarefas do cotidiano, onde muitas vezes estive ausente e, no final, ser meu motorista e minha "bengala". Ao meu "querido genro", Rodolpho, pelo apoio técnico que foi de extra importância na realização deste trabalho.

Ao meu netinho e Amado Vítor, que é um exemplo de vitória. Mesmo a vida falando não, ele não desiste.

Aos amigos de caminhada, Natália, Matheus, Deiverson, Lucas e Rafael, pelo companheirismo nestes últimos anos, principalmente na pandemia. Os "*helps*" de vocês foram muito importantes no desenrolar não só deste estudo como também da graduação em si. A equipe do Projeto JF Salvando todos, de onde surgiu a inspiração do tema desta pesquisa.

Aos funcionários do Instituto de Ciências Exatas, principalmente do Departamento de Estatística, pelo zelo com o local, pelo cuidado com nossa segurança e por pequenos gestos que propiciaram um grande conforto nas horas que passei por ali.

Enfim, só agradeço. Infinitamente agradeço. De coração.

“Tudo posso naquele que fortalece.”
(Filip 4, 13)

RESUMO

As estimativas de pesquisas são frequentemente afetadas por erros amostrais e não amostrais. Combinar dados de mais de uma pesquisa pode ser benéfico para a produção de estimadores mais consistentes para a população de interesse, uma vez que amplia o tamanho total da amostra, reduzindo erros de amostragem. Pode-se combinar não apenas levantamentos separados, mas também amostras contínuas do mesmo levantamento, bem como dados de painéis sobrepostos em um levantamento em painel. Neste contexto, o objetivo deste estudo foi combinar dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) coletados em 2021 e 2022, com o intuito de reduzir os erros de amostragem das estimativas da população brasileira. Para tanto, usou-se o método desenvolvido por Dzikiti, Vieira e Girdler-Brown (2023), para buscar estimadores combinados a partir de duas ou mais pesquisas, usando como ponderadores o inverso do coeficiente de variação e o inverso do efeito do plano amostral ampliado. Também se usou o inverso do tamanho amostral e o inverso da variância proposto por Fox (2010) nos seus estudos sobre metanálise. Para aplicação dos métodos estudados, foi usada a variável quantitativa "rendimento médio no trabalho principal" (VD4020). Conclui-se que, ao combinar as pesquisas, houve melhora na precisão das estimativas com erros padrão menores do que seria observado quando somente uma pesquisa é utilizada.

Palavras-chave: Abordagem separada e combinada, metanálise, amostragem complexa.

ABSTRACT

Survey estimates are often affected by sampling and non-sampling errors. Therefore, combining data from more than one survey can be beneficial for producing more consistent estimators for the population of interest, as it expands the total sample size, reducing sampling errors. You can combine not only separate surveys, but also continuous samples from the same survey, as well as overlapping panel data into a panel survey. In this context, the objective of this study was to combine data from the the *Pesquisa Nacional por Amostra de Domicílios Contínua* (PNADC) collected in 2021 and 2022, with the aim of reducing sampling errors in the estimates of the brazilian population. To this end, the method developed by Dzikiti, Vieira and Girdler-Brown (2023) was used to seek combined estimators from two or more surveys, using as weights the inverse of the coefficient of variation and the inverse of the effect of the sampling plan enlarged. We also used the inverse of the sample size and the inverse of the variance proposed by Fox (2010) in her studies on meta-analysis. To apply the studied methods, the quantitative variable "*rendimento efetivo no trabalho principal*" (VD4020) was used. It is concluded that, when combining the surveys, there was an improvement in the accuracy of the estimates with standard errors smaller than what would be observed when only one survey is used.

Keywords: Separate and combined approach, meta-analysis, complex sampling**.**

Sumário

1. INTRODUÇÃO

As estimativas de pesquisas amostrais são frequentemente afetadas por erros amostrais e não amostrais. Dados ausentes, erro de cobertura e erro de medição ou resposta são enumerados como exemplos de erros não amostrais, podendo ser difíceis de serem avaliados e corrigidos quando se usa informações de um único estudo. Assim, combinar dados de mais de uma pesquisa pode ser benéfico para a produção de estimativas para a população de interesse. Schenker e Raghunathan (2007) discutem esses tipos de erros não amostrais e como a combinação de dados de várias pesquisas pode ajudar a "diminuir" o seu efeito nas estimativas finais.

Existem várias razões pelas quais os analistas gostariam de combinar os dados de duas ou mais pesquisas. Um dos principais motivos é que os tamanhos de amostra para o fenômeno em estudo podem ser pequenos em cada uma das fontes de dados, devido a cada pesquisa ter um tamanho de amostra pequeno ou devido ao domínio de interesse ser raro na população visada por cada uma delas.

Ter tamanhos de amostra pequenos em cada uma das fontes de dados não é a única razão para combinar os dados de duas ou mais pesquisas. Também, um analista pode reunir os dados de pesquisas periódicas sobre o mesmo tópico para estimar a mudança. Ou, nos casos em que pode haver deficiências de estrutura, pode-se usar a combinação de pesquisas com variáveis semelhantes usando métodos de estrutura múltipla (*multiple frame methods*) para melhorar a cobertura (ROBERTS e BINDER, 2009).

A combinação de amostras para aumentar o número de observações é usada não apenas no caso de levantamentos separados, mas também para combinar amostras contínuas do mesmo levantamento e para combinar dados de painéis sobrepostos em um estudo longitudinal, por exemplo. Em todos os casos, espera-se que o aumento do tamanho total da amostra leve a erros de amostragem menores quando em comparação com pesquisas individuais, ou seja, às vezes é necessário combinar dados de diferentes pesquisas para aumentar o tamanho amostral e a precisão das estimativas para subpopulações de interesse ou para melhorar a cobertura, quando em comparação com pesquisas individuais (DZIKITI, 2019).

Com a crescente disponibilidade de dados de mais de um estudo contendo as mesmas variáveis ou variáveis semelhantes, o fato de combinar as diferentes pesquisas com o objetivo de melhorar as estimativas tem merecido uma maior atenção na literatura da Estatística e especificamente da Amostragem. Roberts e Binder (2009) apontam ser razoável pensar que normalmente se deve ser capaz de melhorar a estimativa de uma quantidade de interesse (com relação à exatidão ou precisão) combinando amostras, desde que uma abordagem apropriada seja usada para calcular a nova estimativa.

No entanto, a definição do método a ser usado para esta finalidade nem sempre é claro. Embora haja uma literatura substancial sobre estudos que agrupam dados de pesquisas, ainda não está evidente quais são as metodologias e métodos de amostragem mais eficientes para agrupar dados de estudos diferentes. Por exemplo, é importante saber se as estimativas das pesquisas envolvidas devem receber pesos iguais no cálculo da estimativa combinada ou não. Se eles não recebem igual importância, então deve ficar claro como eles devem ser ponderados e por quê (DZIKITI, 2019).

Para que pesquisas sejam comparadas ou combinadas corretamente, a definição de conceitos, variáveis de interesse, populações, os métodos de medição e a análise substantiva devem ser semelhantes. A suposição chave na combinação de pesquisas é que as populações-alvo das pesquisas sejam equivalentes (FOX, 2010).

Dzikiti (2019) e Dzikiti, Vieira e Girdler-Brown (2023), em seus estudos, propõem métodos para combinar dados de pesquisa e os avaliam por meio de simulações no contexto de amostragem aleatória simples, amostragem aleatória estratificada e amostragem por conglomerados em dois estágios.

Neste contexto, o objetivo do presente estudo, é aplicar os métodos desenvolvidos por Dzikiti, Vieira e Girdler-Brown (2023) para a combinação de duas edições da pesquisa do IBGE – Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC), dos anos de 2021 e 2022, para redução dos erros de amostragem das estimativas na população de interesse, ou seja, os habitantes do Brasil.

Na Seção 2, há uma breve descrição dos conceitos de Amostragem e de seus principais modelos (probabilísticos) e dos erros ligados à obtenção das amostras. Há também, uma breve introdução sobre os métodos de combinação de pesquisas. A Seção 3 começa com definições das duas abordagens usuais para estimação a partir de dados combinados de várias pesquisas e, em seguida, com uma discussão sobre a escolha dos

fatores de redimensionamento (pesos) como o coeficiente de variação (cv) do estimador e o efeito do plano amostral ampliado (meff). Uma aplicação da combinação dos dados de duas pesquisas baseado nos resultados do trabalho de Dzikiti, Vieira e Girdler-Brown (2023) é descrita na Seção 4. E o documento termina na seção 5 com a discussão dos resultados obtidos.

2. REFERENCIAL TEÓRICO

A amostragem é um procedimento para selecionar elementos de uma população (amostra), de forma que seja possível fazer análises e chegar a conclusões sobre a população. De acordo com Bolfarine e Bussab (2005), o objetivo da amostragem é fazer afirmações sobre uma população, baseando no resultado (informação) de uma amostra, em que essa é obtida a partir de uma população bem definida.

Para a escolha de um método de amostragem deve-se levar em conta alguns critérios como o tipo de pesquisa, a acessibilidade e disponibilidade dos elementos da população, a disponibilidade de tempo, os recursos financeiros e humanos, dentre outros.

Existem diversas maneiras para se selecionar uma amostra de uma população que podem ser classificadas em métodos probabilísticos, em que cada elemento da população possui uma probabilidade não-nula de ser selecionado para compor a amostra, fazendo uso de mecanismos aleatórios de seleção. E não probabilísticos, em que a seleção da amostra depende do julgamento do pesquisador, podendo haver uma escolha deliberada dos elementos para compor a amostra, cujos mecanismos não aleatórios de seleção são evidenciados.

De acordo com Bolfarine e Bussab (2005), os métodos de amostragem probabilística podem ser classificados em Amostragem Aleatória Simples (AAS) com e sem reposição, Amostragem Sistemática, Amostragem Estratificada, Amostragem por conglomerados, dentre outros.

2.1 AMOSTRA ALEATÓRIA SIMPLES

A Amostragem Aleatória Simples (AAS) é um método comumente utilizado para selecionar uma amostra de indivíduos de forma aleatória e imparcial. Neste tipo de

amostragem, cada membro da população tem uma chance igual de ser selecionado para compor a amostra. A classificação desse método em relação à reposição pode ser dividida em dois tipos principais: Amostragem Aleatória Simples com reposição e Amostragem Aleatória Simples sem reposição.

A Amostragem Aleatória Simples com reposição (AASc) permite que o mesmo elemento da população possa ser amostrado mais de uma vez. A seleção da amostra por AASc resulta na atribuição da mesma probabilidade de seleção para todas as unidades da população, ou seja, se a população tem um tamanho **N**, cada elemento tem a mesma probabilidade de seleção igual a **1/N**. É o método mais simples, porém o mais importante, pois garante a independência entre as observações, o que é um facilitador na determinação das propriedades dos estimadores dos parâmetros.

Já na Amostragem Aleatória Simples sem reposição (AASs), cada elemento pode ser amostrado uma única vez. É um procedimento de seleção que garante que todas as amostras de tamanho n têm a mesma probabilidade de serem selecionadas (BOLFARINE e BUSSAB, 2005). À medida que os elementos são selecionados, a probabilidade de escolha para os elementos restantes muda, pois os elementos já escolhidos são removidos da população. Na AASs, a independência é garantida, pois cada elemento é único e não é substituído na população após a seleção, o que significa que a escolha de um elemento não afeta a escolha dos outros elementos subsequentes.

A precisão de qualquer estimativa feita com base em uma amostra depende tanto do método pelo qual a estimativa é calculada a partir dos dados da amostra quanto do plano de amostragem. Para amostragem aleatória simples, tem-se $\overline{Y}$ como estimador da média da população.

O estimador da média amostral ($\bar{y}$) é usado para estimar a média da população. É calculado como a média dos valores observados na amostra:

$$\bar{y} = (y_1 + y_2 + \cdots + y_n)/n,$$

em que:

$\bar{y}$ é o estimador da média amostral,

$(y_1 + y_2 + \cdots + y_n)$ são os valores observados na amostra e

n é o tamanho da amostra.

Segundo Cochram (1999), a média amostral é um estimador não viesado de $\overline{Y}$. Logo:

$$E(\bar{y}) = \overline{Y}.$$

E a variância do estimador da média é dada por

$$V(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{S_y^2}{n},$$

em que S_y^2 é a variância populacional, N é o tamanho populacional e n é o tamanho amostral.

E o estimador da variância é dado por:

$$\hat{V}(\bar{y}) = \left(\frac{N-n}{N}\right)\frac{s^2}{n},$$

em que s^2 é a variância amostral, N é o tamanho populacional e n é o tamanho amostral.

2.2 AMOSTRAGEM SISTEMÁTICA

Na Amostragem Sistemática (AS), os elementos da população são ordenados de alguma forma – alfabeticamente ou através de algum outro método (filas, por exemplo). Um ponto de partida aleatório é sorteado, e então cada k-ésimo membro da população é selecionado para a amostra, ou seja, o procedimento amostral consiste em selecionar cada k-ésima unidade de um cadastro, começando de uma partida aleatória, em que as unidades são selecionadas sem reposição e com a mesma probabilidade de serem selecionadas (1/k) (COCHRAN,1999).

Seja Y o total para a amostra s_r, r = 1, ..., n

$Y = \sum_{i=1}^{n} y_i$, em que n é o tamanho da amostra s_r.

Assim a média amostral é dada por

$$\bar{y} = \frac{Y}{n}.$$

Pode-se mostrar que E ($\bar{y}$) ≠ $\bar{Y}$ para a AS (ver Cochran, 1999, pg. 207). E que não se pode estimar diretamente variâncias de estimadores usando este método. A adoção de métodos de replicação é uma alternativa.

2.3 AMOSTRAGEM ESTRATIFICADA

Na Amostragem Estratificada (AE), a população é inicialmente dividida em subgrupos (estratos) que são mutualmente exclusivos e exaustivos, de acordo com características conhecidas. As unidades de cada estrato são selecionadas de forma independente. Os subgrupos geralmente são internamente mais homogêneos que a população como um todo, o que pode proporcionar uma redução de erro amostral. Podem ser adotados planos amostrais alternativos nos diferentes estratos (BOLFARINE e BUSSAB, 2005).

Se dentro de cada estrato, uma amostra aleatória simples for selecionada, tem-se a Amostragem Estratificada Simples (AES) e o tamanho da amostra pode ser proporcional ao tamanho do estrato. Pode-se identificar as unidades populacionais usando dois rótulos: h (h =1, ..., H), que indica o estrato a que pertence e i (i = 1, ..., N_h), que indica a unidade dentro do estrato.

De acordo com Kish (1995), a média populacional é igual à soma das médias dos H estratos $\bar{Y}_h$ multiplicado pelo peso proporcional W_h, onde $\sum W_h = 1$. A média amostral ponderada é dada por

$$\bar{y} = \sum W_h \, \bar{y}_h.$$

A média amostral estimada é dada por

$$\bar{y}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h} \qquad \text{e} \qquad E(\bar{y}_h) = \bar{Y}.$$

A variância desta média é

$$s_h^2 = \frac{(y_{hi} - \bar{y}_h)^2}{n_h - 1} \quad \text{e} \quad E(s_h^2) = S_h^2.$$

O estimador da variância é dado por

$$\widehat{V}(\bar{y}) = \sum_{h=1}^{H} w_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2.$$

De acordo com Bolfarine e Bussab (2005), os planos amostrais apresentados acima consistem em sorteios de unidades elementares diretamente da população ou de estratos desta mesma população.

2.4 AMOSTRAGEM POR CONGLOMERADOS

A amostragem por conglomerados é um método de amostragem estatística no qual a população é dividida em grupos ou conglomerados, e em seguida, uma seleção de conglomerados é feita para compor a amostra. Ao contrário da AAS, em que cada indivíduo da população tem a mesma chance de ser selecionado, na AC, é a unidade de conglomerados que é selecionada aleatoriamente (BOLFARINE e BUSSAB, 2005).

Os conglomerados são formados de modo que cada um represente de forma adequada a heterogeneidade da população em estudo. Uma vez que os conglomerados são selecionados aleatoriamente, todas as unidades dentro de cada conglomerado são incluídas na amostra. Isso significa que na Amostragem por conglomerados é um processo de dois ou mais estágios.

De acordo com Bolfarine e Bussab (2005), a Amostragem por conglomerados pode ser em Um Estágio ou Múltiplos estágios.

Na amostragem por conglomerados em Um Estágio, uma amostra de conglomerados é selecionada de acordo com um plano amostral qualquer e todos os elementos pertencentes aos conglomerados selecionados compõem a amostra.

Na Amostragem por Conglomerados em Múltiplos Estágios, os conglomerados são homogêneos e se torna menos aconselhável a seleção de todos os seus elementos. No caso, é feito um sorteio de elementos dos conglomerados selecionados (BOLFARINE e BUSSAB, 2005).

Quando se tem a Amostragem por Conglomerados em dois estágios, seleciona-se no primeiro estágio, conglomerados (unidades primárias de amostragem – UPAs) segundo algum plano amostral. De cada conglomerado, sorteia-se elementos (unidades secundárias de amostragem – USAs) através do mesmo ou de outro plano amostral (BOLFARINE e BUSSAB, 2005).

Para a Amostragem por Conglomerados em três estágios, a primeira etapa consiste na amostra de UPAs selecionadas. Na segunda etapa, tem-se uma amostra de USAs selecionada de cada uma das UPAs selecionadas anteriormente. E por fim, a terceira etapa consiste de amostra de unidades elementares selecionadas de cada uma das USAs selecionadas.

Ao se selecionar i conglomerados por AAS, dentre os N existentes e pesquisar todas as unidades nos conglomerados selecionados, tem-se a Amostragem por Conglomerados Simples (ACS), em que os valores amostrais é dado por

$$y_{ij} \ \forall \, j=1,\ldots,M_i \quad e \quad i=1,\ldots n \ ,$$

em que M_i é o número de unidades no conglomerado i.

A estimação da média por unidade elementar, $\bar{y}$ - estimador "natural", é dada por

$$\bar{y}_N = \frac{\hat{Y}}{M_0} = \frac{N}{M_0}\sum_{i=1}^{n} y_i/n = \frac{\bar{y}_c}{\bar{M}},$$

em que M_0 representa todas as unidades da população. E $\bar{y}_N$ é um estimador não viciado.

A variância do estimador natural é dada por

$$V(\bar{y}_N) = \frac{N^2}{{M_0}^2}\frac{1-f}{n}S_c^2 = \frac{1}{\bar{M}^2}\frac{1-f}{n}S_c^2.$$

E o estimador da variância do estimador natural é dado por

$$\hat{V}(\bar{y}_N) = \frac{1}{\bar{M}^2}\frac{1-f}{n}s_c^2.$$

2.5 AMOSTRAGEM COM PROBABILIDADES PROPORCIONAIS AO TAMANHO

Todos os planos amostrais abordados acima atribuem a todas as amostras possíveis de serem selecionadas a mesma probabilidade de seleção. As unidades de amostragem têm variação de tamanho que, ao ser ignorada, pode resultar em desenhos ineficientes. Quando essa variação for grande, é conveniente usar a Amostragem com Probabilidades Proporcionais ao Tamanho (PPT), que é um método de amostragem utilizado para selecionar elementos de uma população de acordo com sua magnitude ou

tamanho relativo. Esse método é frequentemente utilizado quando a população possui unidades de diferentes tamanhos.

Sejam δ_i as variáveis indicadoras de inclusão na amostra s, para todo i ∈ U. Para um plano amostral p(s) qualquer, tem-se que

$$E_p(\delta_i) = \pi_i\,, \qquad E_p(\delta_i\delta_j) = \pi_{ij}\,,$$

$$V_p(\delta_i) = \pi_i\,(1 - \pi_i\,) \qquad \text{e} \qquad COV_p(\delta_i;\,\delta_i) = \pi_{ij} - \pi_i\,\pi_j\,.$$

Seja o total populacional, $Y = \sum_U Y_i$, o parâmetro alvo. A estimação linear de Y é dada por

$\widehat{Y} = \sum_S w_i y_i = \sum_U w_i y_i \delta_i$, em que w_i é o peso da unidade i.

Se o estimador acima não for viciado, tem-se que

$$E_p(\widehat{Y}) = Y \;\Rightarrow\; \sum_U w_i y_i E_p(\delta_i) = \sum_U y_i \;\Rightarrow\; \sum_U w_i \pi_i\, y_i = \sum_U y_i\,.$$

Esta relação só será válida para quaisquer valores y_i da variável de pesquisa caso $w_i * \pi_i = 1$, para qualquer i ∈ U. E a condição para que o estimador de total seja não viciado é que os pesos das unidades na amostra seja iguais ao inverso das respectivas probabilidades de inclusão na amostra, ou seja,

$$w_i = \pi_i{}^{-1}, \quad \forall\, i \in U.$$

Logo, o estimador não viciado do total fica dado por

$$\widehat{Y} = \sum_S \frac{y_i}{\pi_i}.$$

Como o tamanho da população N é conhecido, o estimador “natural” da média populacional é dado por

$$\overline{Y}_{HT} = {}^{\widehat{Y}_{HT}}\!/_{N} = \sum_S \frac{y_i}{N\pi_i} = \frac{1}{N}\sum_S \pi_i^{-1}\, y_i = \sum_S w_i^{HT}\, y_i.$$

2.6 ERROS AMOSTRAIS E NÃO AMOSTRAIS

Levantamentos amostrais podem ser acometidos por diferentes tipos de erros que podem ser classificados como amostrais ou não amostrais. Os primeiros podem ser entendidos como sendo resultantes da diferença entre a estimativa produzida a partir da

amostra e o verdadeiro parâmetro da população. Não importa quão bem a amostra seja coletada, os erros amostrais irão ocorrer. Cada vez que uma amostra aleatória for retirada de uma população, um resultado diferente será observado. Isso ocorre porque as amostras são aleatórias.

Já os erros não amostrais ocorrem quando a coleta dos dados é acometida por erros, seja devido ao uso de algum instrumento de medida defeituoso, ou anotações erradas, dentre outros fatores. Dados faltantes, erro de cobertura e erro de medição ou resposta são enumerados como exemplos de erros não amostrais.

Bolfarine e Bussab (2005) diz que todo levantamento amostral ou não, está sujeito a produzir diferenças entre o parâmetro populacional θ, de interesse, e o valor de $\hat{\theta}$ empregado para estimá-lo. A diferença $\hat{\theta} - \theta$ é considerado como erro de pesquisa.

Cochran (1999) mostra que para comparar um estimador viesado com um não viesado, ou dois estimadores com diferentes tamanhos de viés, um critério normalmente usado é o erro quadrático médio (EQM). Formalmente,

$$\text{EQM}(\hat{\theta}) = (\text{var}(\hat{\theta})) + (\text{viés})^2.$$

A fórmula geral para calcular o EQM é a seguinte

$$\text{EQM} = (1/p) * \Sigma (y_i - \hat{y}_i)^2, \text{ em que}$$

EQM: Erro Quadrático Médio,

p: número de amostras,

y_i: valor real do i-ésimo dado, e

$\hat{y}_i$: valor previsto pelo modelo para o i-ésimo dado.

Quanto menor o valor do EQM, melhor é a qualidade do estimador. Pode-se dizer que o EQM penaliza de forma mais significativa erros maiores, devido ao cálculo dos quadrados das diferenças, o que significa que valores discrepantes têm um impacto mais significativo no EQM.

2.7 COMBINAÇÃO DE PESQUISAS

A fim de minimizar os erros de amostragem, alguns estudos têm usado uma técnica de combinação de informações de mais de uma pesquisa, em que se pode tirar proveito dos pontos fortes de pesquisas diferentes e usar em um estudo para fornecer

informações que faltam em outro, ajustando os erros não amostrais. Além disso, as estimativas combinadas resultantes podem ter níveis de erro de amostragem menores, com estimadores mais consistentes. Assim, combinar informações de várias pesquisas pode fornecer informações aprimoradas das estimativas de interesse (SCHENKER e RAGHUNATHAN, 2007).

Em todos os casos, espera-se que o aumento do tamanho total da amostra leve a erros de amostragem reduzidos, ou seja, às vezes é necessário combinar dados de diferentes pesquisas para aumentar o tamanho amostral e a precisão das estimativas para subpopulações de interesse ou para melhorar a cobertura, quando em comparação com pesquisas individuais (ROBERTS e BINDER, 2009).

Dzikiti, Vieira e Girdler-Brown (2023), dizem que, embora haja uma literatura substancial sobre estudos que agrupam dados de pesquisas, ainda não está claro quais as metodologias e desenhos de amostragem são mais eficientes para agrupar dados de pesquisas diferentes. É importante saber se as estimativas das pesquisas envolvidas devem receber pesos iguais no cálculo da estimativa combinada ou não. Se os pesos não recebem igual importância, então deve ficar claro como eles devem ser ponderados e o porquê.

Para os estudos que serão combinados, é importante que a definição de conceitos, variáveis de interesse e populações devem ser feitos corretamente e os métodos de medição e a análise descritiva devem ser semelhantes. A suposição chave na combinação de pesquisas é que as populações-alvo das pesquisas sejam equivalentes (Fox, 2010). Sempre que possível, testes de homogeneidade podem ser usados para verificar essa equivalência.

Roberts e Binder (2009), relatam que combinar os dados de mais de uma fonte levanta uma série de questões que precisam ser abordadas antes que decisões razoáveis possam ser tomadas e como as estimativas podem ser realizadas usando as diferentes fontes. A primeira delas é a comparabilidade das informações obtidas nas diferentes pesquisas. Schenker et al (2002) e Schenker e Raghunathan (2007) discutem uma série de fontes potenciais de incomparabilidade que podem afetar se as variáveis registradas em diferentes pesquisas estão realmente medindo as mesmas quantidades: diferenças nos tipos de respondentes e/ou nas fontes de informação dos respondentes, diferenças nos

modos de entrevista, diferenças nos contextos de pesquisa, diferenças nos desenhos amostrais e diferenças nas perguntas da pesquisa.

No entanto, uma importante questão de comparabilidade está relacionada a como as populações-alvo das fontes de dados se comparam: se são semelhantes tanto para o grupo-alvo quanto para o tempo, se os grupos-alvo são semelhantes, mas os tempos diferem (que é o caso mais comum) ou se eles diferem substancialmente em relação ao grupo-alvo e ao tempo (ROBERTS e BINDER, 2009).

Existem duas abordagens principais que podem ser usadas ao combinar pesquisas, que são as abordagens combinada e separada. Ambas abordagens têm suas vantagens e desvantagens.

Conceituando as duas abordagens, pode-se dizer que na abordagem combinada, os registros individuais das pesquisas são mesclados, resultando em um aumento do tamanho e do poder da amostra. Os pesos do desenho amostral da pesquisa original podem ser modificados e as estimativas podem ser calculadas com base nos novos pesos e na amostra combinada. Só é possível realizar uma análise combinada quando os dados das pesquisas individuais estão disponíveis. Além disso, uma vez calculada uma estimativa combinada, não há necessidade de que volte para os conjuntos de dados individuais da pesquisa. No entanto, a abordagem conjunta requer mais conhecimento técnico na manipulação de arquivos de dados quando comparado à abordagem separada (DZIKITI, VIEIRA e GIRDLER-BROWN, 2023).

Por outro lado, na abordagem separada, uma estimativa é obtida de cada pesquisa separadamente e, então, usadas para calcular uma estimativa combinada. O método mais comum é usar alguma combinação linear das estimativas separadas para formar o estimador combinado. A combinação linear escolhida pode depender se a quantidade de interesse é descritiva ou analítica. A combinação linear também pode depender se as estimativas de pesquisa separadas são independentes e se é possível obter uma redução adequada nas variações da estimativa geral para as quantidades de interesse mais importantes.

A combinação de pesquisas está intimamente relacionada à meta-análise, que visa melhorar a precisão da estimativa de um tamanho de efeito por meio do agrupamento de informações entre os estudos.

Segundo Deekes et al (2022), a meta-análise é tipicamente um processo de dois estágios. Primeiramente, uma estatística resumida é calculada para cada estudo, para descrever o efeito da intervenção observado da mesma forma para todos os estudos.

Na segunda etapa, uma estimativa resumida (combinada) do efeito da intervenção é calculada como uma média ponderada dos efeitos da intervenção estimados nos estudos individuais. Uma média ponderada é definida como

$$\text{Média ponderada} = \frac{\sum_{i=1}^{n} Y_i w_i}{\sum_{i=1}^{n} w_i},$$

em que Y_i é o efeito da intervenção estimado no i-ésimo estudo e w_i é o peso dado ao i-ésimo estudo. Observe que, se todos os pesos forem iguais, a média ponderada é igual ao efeito médio da intervenção. Quanto maior o peso dado ao i-ésimo estudo, mas ele contribuirá para a média ponderada.

3. METODOLOGIA

Existem várias metodologias para combinar múltiplas amostras que podem ser divididas entre métodos que usam análises separadas para as amostras e métodos que combinam os dados e realizam uma análise combinada, ou seja, métodos que calculam estimativas para cada pesquisa e depois as combinam e métodos que combinam os dados de todas as pesquisas como um grande banco de dados antes de calcular a estimativa combinada.

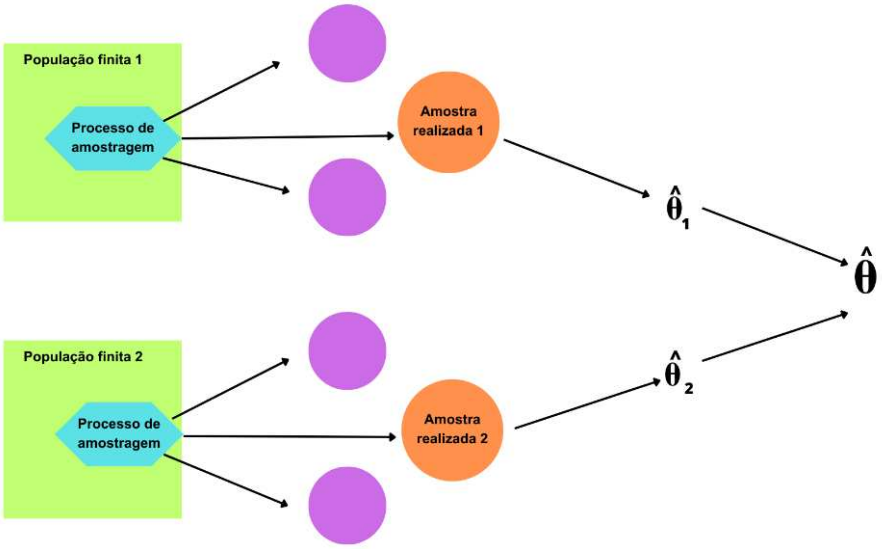
Roberts e Binder (2009), relatam que na abordagem separada, uma estimativa é obtida de cada pesquisa de forma distinta e, consequentemente, o estimador geral é uma função das estimativas separadas. O método mais comum é usar alguma combinação linear das estimativas separadas para formar um estimador geral onde uma estimativa combinada é calculada em função das estimativas separadas, geralmente baseadas em uma combinação linear.

Por exemplo, no caso de duas pesquisas,

$$\hat{\theta}_c = \alpha\hat{\theta}_1 + (1 - \alpha)\hat{\theta}_2,$$

em que $\hat{\theta}_c$ é o estimador combinado do parâmetro alvo $\hat{\theta}$, $\hat{\theta}_d$ é o estimador com base nos dados da pesquisa d, com d = 1, 2 e α é um peso alocado para a pesquisa.

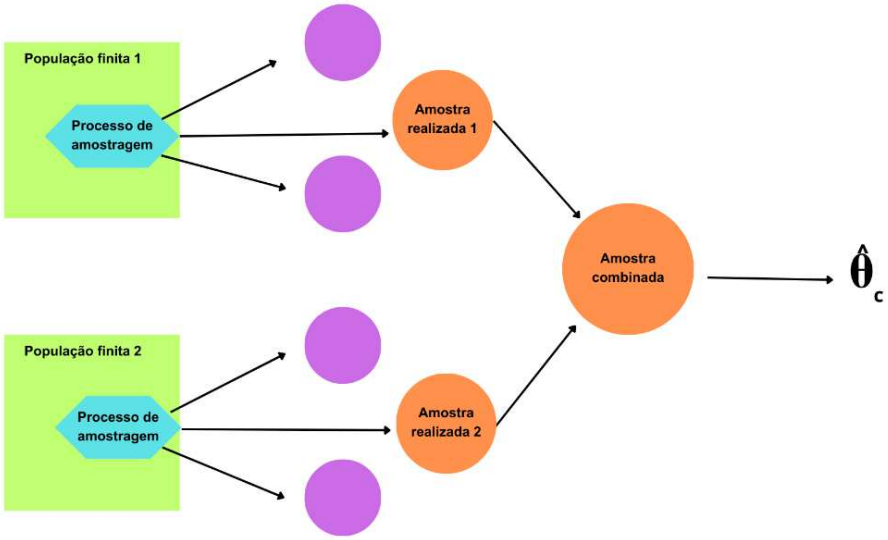
Figura 1 - Abordagem separada



Fonte: Robert e Binder (2009)

Por outro lado, na abordagem combinada, os registros individuais de todas as pesquisas são combinados, os pesos originais podem ser modificados e a estimativa é baseada na amostra combinada usando os novos pesos e utilizando técnicas apropriadas para uma única amostra. Normalmente, para as observações em cada pesquisa individual, os pesos modificados são proporcionais aos pesos originais (ROBERTS e BINDER, 2009).

Figura 2 - Abordagem combinada



Fonte: Robert e Binder (2009)

Dzikiti, Vieira e Girdler-Brown (2023) dizem que a combinação de pesquisas está intimamente relacionada à meta-análise, que busca melhorar a precisão da estimativa por meio do agrupamento de informações provenientes de diferentes estudos. Geralmente, a meta-análise usa o inverso da variância da estimativa calculada para cada estudo individual para ponderar as estimativas dos diferentes estudos no cálculo do tamanho de efeito combinado. Desta maneira, estudos menos precisos recebem menor peso na estimativa do tamanho do efeito combinado.

A escolha dos fatores de redimensionamento pode depender de critérios semelhantes aos usados para escolher uma combinação linear na abordagem separada. Fox (2010), com o objetivo de dar uma estrutura para a meta análise, comparou o tamanho da amostra, a variância inversa e os métodos de ponderação da média para estimativas de pesquisas combinadas. Maheswaran et al (2015) usaram o inverso da variância ao combinarem pesquisas por meio de regressão ponderada em estudo transversal de saúde autorrelatada e desigualdades socioeconômicas na Inglaterra. Scho (2017), integrando estimativas da média para duas pesquisas, buscou modelos para erro de medição (erros não amostrais). Além disso, Kish (1995), sugeriu o uso de pesos ou tamanhos de amostra iguais para ponderar amostras ao combinar pesquisas.

Alguns autores tem usado uma média direta da estatística descritiva obtida de pesquisas individuais como o estimador de pesquisas combinadas que pode ser calculado como

$$\hat{\theta}_c = \frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d \ ,$$

em que $\hat{\theta}_c$ é o estimador combinado e $\hat{\theta}_d$ é o estimador obtido de pesquisas individuais d = 1, ..., D.

Porém, se houver agrupamento, como costuma acontecer com pesquisas por amostragem complexa, isso pode nem sempre ser verdade. Usar o tamanho efetivo da amostra para ponderação pode ser uma forma de se lidar com esse problema. Vale ressaltar que o tamanho efetivo da amostra é usado para medir a representatividade de uma amostra em relação à população alvo. Em amostragens complexas, o tamanho efetivo é usado para ajustar a variância dos estimadores e levar em consideração o plano amostral utilizado.

Kish, em 1965, introduziu o termo *Design Effectt* (*deff*), ou seja, o efeito do desenho amostral. Ele calculou o *deff* como

$$\mathrm{deff}_{\mathrm{kish}}\left(\hat{\theta}\right) = \frac{\mathrm{Var}_{\mathrm{V}}\left(\hat{\theta}\right)}{\mathrm{Var}_{\mathrm{AAS}}\left(\hat{\theta}\right)},$$

em que $\mathrm{Var}_{\mathrm{V}}\left(\hat{\theta}\right)$ é a verdadeira variância de $\hat{\theta}$ que considera o verdadeiro esquema de amostragem usado para a seleção da amostra, e $\mathrm{Var}_{\mathrm{AAS}}\left(\hat{\theta}\right)$ é a variância hipotética de $\hat{\theta}$ quando se supõe que a amostra foi selecionada por amostragem aleatória simples com reposição.

Skinner (1985) propôs o efeito de má especificação, *misspecification effect* (*meff*), conhecido na literatura brasileira como o efeito do plano amostral ampliado que é mais adequado para análises inferenciais do que o efeito de desenho amostral (*deff*). O *meff* visa medir os efeitos incorretos de especificação tanto do esquema de amostragem quanto do modelo considerado. O *meff* é definido de forma semelhante ao *deff* com o mesmo numerador, mas com o denominador que consiste na esperança de um estimador de variância que ignora o desenho complexo.

Dzikiti, Vieira e Girdler-Brown (2023) consideraram a abordagem separada e exploraram a importância de medidas alternativas para as diferentes pesquisas envolvidas na estimativa combinada como o coeficiente de variação (cv), a assimetria, a curtose e Teste D'Agostino-Pearson usadas como estratégia de ponderação quando combinando pesquisas, sob a abordagem separada. Além disso, a eficiência de três desenhos amostrais foi considerada para as estimativas de pesquisas que estão sendo combinadas – Amostragem Aleatória Simples sem reposição, Amostragem Estratificada e Amostragem por conglomerados em dois estágios. No caso do presente estudo, usou-se apenas o coeficiente de variação (cv) do estimador e o efeito do plano amostral ampliado (*meff*).

O coeficiente de variação (cv) é estimado pela razão do erro padrão do estimador e o estimador pontual, o que justifica a independência da escala de medida (KISH,1995), dado por

$$\mathrm{cv} = \frac{\mathrm{ep}\left(\hat{\theta}\right)}{\hat{\theta}} * 100\%.$$

Ao combinar estimativas de diferentes pesquisas, deve-se considerar os valores de cv como uma medida de variação mais razoável para os estimadores do que a variância, ou seja, é desejável dar mais peso a levantamentos com maior precisão. Isso ocorre porque

o CV leva em conta a dispersão relativa em relação à media, o que pode refletir a precisão dos estimadores. Vale ressaltar que a variância é uma medida de dispersão absoluta dos valores em torno da média, enquanto o coeficiente de variação é uma medida relativa da variação, permitindo a comparação da variação entre diferentes estimadores ou conjunto de dados, independentemente das unidades de medida. Assim, Dzikiti, Vieira e Girdler-Brown (2023) propõem que ao calcular estimativas combinadas, deve-se ponderar pelo inverso do seu cv, ou seja,

$$\hat{\theta}_c = \sum_{d=1}^{D} \left(\frac{\frac{1}{cv(\hat{\theta}_d)} * \hat{\theta}_d}{\sum_{d=1}^{D} \frac{1}{cv(\hat{\theta}_d)}} \right).$$

Dzikiti, Vieira e Girdler-Brown (2023) também propuseram que o inverso do efeito do plano amostral ampliado (*meff*) pode ser usado como peso quando se combina estimativas de pesquisas de amostragem complexas.

De acordo com Skinner (1985), o *meff* é definido como

$$\text{meff}(\hat{\theta}, \text{Var}_0) = \frac{\text{Var}_v(\hat{\theta})}{E_v[\text{Var}_0(\hat{\theta})]},$$

em que $\text{Var}_v(\hat{\theta})$ é a verdadeira variância de $\hat{\theta}$, considerando o desenho amostral adotado para selecionar a amostra enquanto $\text{Var}_0(\hat{\theta})$ é o estimador de variância assumindo AASs. As medidas de meff quanto à $\text{Var}_0(\hat{\theta})$ superestima ou subestima a $\text{Var}_v(\hat{\theta})$ e pode ser estimado por

$$\widehat{\text{meff}}\left(\hat{\theta}, \text{Var}_0\right) = \frac{\text{Var}_v(\hat{\theta})}{\text{Var}_0(\hat{\theta})},$$

em que $\text{Var}_v(\hat{\theta})$ é o estimador de $\text{Var}(\hat{\theta})$, de acordo com o desenho amostral. Com isso, Dzikiti, Vieira e Girdler-Brown (2023), relatam que os valores estimados para o meff podem ser interpretados como

i. $\widehat{\text{meff}}\left(\hat{\theta}, \text{Var}_0\right) < 1$, superestima $\text{Var}_v(\hat{\theta})$;
ii. $\widehat{\text{meff}}\left(\hat{\theta}, \text{Var}_0\right) = 1$, sugere estimativa correta de $\text{Var}_v(\hat{\theta})$);
iii. $\widehat{\text{meff}}\left(\hat{\theta}, \text{Var}_0\right) > 1$, → subestima $\text{Var}_v(\hat{\theta})$.

E que o estimador combinado ($\widehat{\theta_c}$), quando as amostras são ponderadas pelo inverso de seu meff, é calculado como

$$\hat{\theta}_c = \sum_{d=1}^{D} \left(\frac{\frac{1}{\widehat{meff}(\hat{\theta}_d)} * \hat{\theta}_d}{\sum_{d=1}^{D} \frac{1}{\widehat{meff}(\hat{\theta}_d)}} \right).$$

Logo, um estimador combinado geral é dado por

$$\hat{\theta}_c = \sum_{d=1}^{D} \left(\frac{w_d * \hat{\theta}_d}{\sum_{d=1}^{D} w_d} \right),$$

em que w_d é o peso relevante que foi alocado para a pesquisa individual.

Para aplicar os desenvolvimentos metodológicos de Dzikiti, Vieira e Girdler-Brown (2023) que usaram de simulações para testar suas hipóteses em relação à combinação de amostras, o presente trabalho combinou duas edições da pesquisa amostral da Pnad contínua, do IBGE, dos anos 2021 e 2022 com o intuito de obter um estimador mais acurado.

4. APLICAÇÕES

É importante nesta parte do trabalho, entender como funciona a PNAD Contínua antes de dar continuidade ao estudo em si.

4.1 PNAD CONTÍNUA

A PNAD Contínua foi implantada em todo o território nacional a partir de janeiro de 2012 e visando produzir indicadores conjunturais relativos à força de trabalho, substituiu as estatísticas sobre mercado de trabalho obtidas pela Pesquisa Mensal de Emprego (PME) e com isso aumentou a cobertura para o território nacional. Também substituiu a Pesquisa Nacional por Amostra de Domicílios (PNAD), o que permitiu a divulgação periódica das informações sobre trabalho e ainda a análise conjuntural do tema (QUINTSLR, 2007). A abrangência nacional é a mesma desagregação geográfica da PNAD: Brasil, Grandes Regiões, Unidades da Federação e Regiões Metropolitanas que incluem os municípios das capitais.

De acordo com as notas técnicas divulgadas pelo IBGE(2018), a PNAD Contínua que, passando a fazer parte, em caráter definitivo, do conjunto de pesquisas correntes do IBGE, tem como principal objetivo a produção de informações contínuas sobre a inserção da população no mercado de trabalho e de características tais como idade, sexo e nível de instrução, bem como permitir o estudo do desenvolvimento

socioeconômico do País através da produção de dados anuais sobre outras formas de trabalho, trabalho infantil, migração, entre outros temas, ou seja, baseando nas características gerais dos moradores.

A pesquisa é realizada por meio de uma amostra probabilística de domicílios, extraída de uma amostra mestra de setores censitários, de forma a garantir a representatividade dos resultados para os diversos níveis geográficos definidos para sua divulgação. De acordo com o IBGE (2018), a amostra mestra é um conjunto de unidades de área selecionadas probabilisticamente de um cadastro mestre, baseado no Censo Demográfico 2010, nas alterações ocorridas na Base Operacional Geográfica e no Cadastro Nacional de Endereços para Fins Estatísticos - CNEFE.

O plano amostral adotado na PNAD Contínua é conglomerado em dois estágios de seleção, com estratificação das unidades primárias de amostragem. A definição de tais unidades levou em consideração o tamanho dos setores censitários, sendo que cada uma delas devia possuir ao menos 60 domicílios particulares permanentes, incluindo os ocupados, os ocupados sem entrevista realizada e os vagos (IBGE, 2018).

No primeiro estágio, são selecionadas as unidades primárias de amostragem (UPA) com probabilidade proporcional ao número de domicílios dentro de cada estrato definido. A estratificação adotada é a definida para todo o Sistema Integrado de Pesquisas Domiciliares - SIPD. A seleção das unidades primárias de amostragem é feita a partir do cadastro mestre, que contém, para cada unidade primária de amostragem, informações sobre a divisão administrativa e algumas características sociodemográficas. As unidades primárias de amostragem que compõem a amostra da PNAD Contínua são as selecionadas para compor a amostra mestra de um trimestre (IBGE, 2018).

De acordo com as notas técnicas do IBGE, 2018, no segundo estágio, que compõem as unidades secundárias de amostragem (USA), são selecionados 14 domicílios particulares permanentes ocupados dentro de cada unidade primária de amostragem da amostra, por amostragem aleatória simples do CNEFE atualizado.

4.2 VARIÁVEIS

Roberts e Binder (2009) ressaltam que uma questão adicional de comparabilidade está relacionada de como as populações alvo das fontes de dados se comparam em relação ao grupo alvo e ao tempo. Ao analisar o primeiro quesito,

verificou-se que as duas amostras são bem semelhantes no que tange ao desenho amostral como também na questão de como foi construído e aplicado o questionário. Em relação ao tempo, buscou-se anos consecutivos (2021 e 2022) da PNAD Contínua.

Um outro pressuposto para a combinação de amostras é que a combinação linear também pode depender se as estimativas de pesquisa separadas são independentes e se é possível obter uma redução adequada nas variações da estimativa geral para as quantidades de interesse mais importantes. Para fins de possibilitar a presente aplicação, considerou-se que as amostras são independentes. Isso se dá porque cada ano de coleta é considerado uma amostra independente pois as informações coletadas em um ano não estão diretamente relacionadas às informações coletadas em anos anteriores ou subsequentes.

As duas pesquisas apresentam dezenas de variáveis que são definidas no documento "Definição das variáveis Derivadas da PNAD Contínua", disponível no site do IBGE cujo endereço está disponível na Referência Bibliográfica desse estudo. Mas para o presente estudo, escolheu a variável "rendimento efetivo no trabalho principal (R$)" a nível Brasil, uma vez que a proposta de estudo de Dzikiti, Vieira e Girdler-Brown (2023), se baseia em variáveis quantitativas. De acordo com Quintslr (2007), o trabalho principal é aquele em que a pessoa efetivamente trabalhou maior número de horas na semana de referência considerada pela pesquisa. Como as categorias de posição na ocupação têm características específicas, as perguntas são diferenciadas para cada uma destas: trabalhador doméstico, empregado, trabalhador não remunerado, conta própria e empregador.

Também se selecionou algumas variáveis categóricas, a fim de se fazer uma análise descritiva dos dados bem como as variáveis de peso ("V1028"), de estrato ("Estrato") e de Unidade Primária de Amostragem ("UPA"). O quadro 1 mostra as variáveis escolhidas para esta análise com suas respectivas descrições.

4.3 ANÁLISE DESCRITIVA DOS DADOS

Antes de prosseguir com a análise principal desse estudo, é importante fazer uma análise descritiva dos dados a fim de entender como se compõe o banco de dados, suas variáveis, dentre outras informações. Para a importação, leitura e salvamento dos

microdados da PNAD Contínua foi usado o pacote *PNADcIBGE*, na interface R Studio, no software livre R, desenvolvido por Braga e Assunção (2020), que fornece um conjunto de ferramentas para baixar, ler e analisar os microdados da PNAD Contínua. Suas funções permitem baixar não apenas os microdados diretamente do site oficial do IBGE como também a documentação e os arquivos necessários para sua leitura e exploração (TROVÃO e SILVA JÚNIOR, 2022). Outras análises foram feitas por meio de pacotes como *tidyverse,* desenvolvido por Wickham et al. (2019), e *survey*, desenvolvido por Lumley (2019) que estão apresentadas no tópico 4.4 "Resultados" deste estudo.

Vale ressaltar que os microdados da PNADC são adequadamente documentados, acompanhados de anexos, notas e dicionários de dados que viabilizam sua interpretação. De acordo com o IBGE (2018), os microdados consistem no menor nível de desagregação dos dados de uma pesquisa, retratando, sob a forma de códigos numéricos, o conteúdo dos questionários, preservado o sigilo estatístico com vistas à não individualização das informações. Os microdados estão no formato ASCII, possibilitando aos usuários especializados, com conhecimento em programação, preferencialmente em softwares estatísticos, a leitura dos dados, o cruzamento em diferentes agregações geográficas, e a elaboração de múltiplas tabulações segundo sua perspectiva pessoal de interesse.

Para a análise exploratória, foram escolhidas, além da variável de interesse desse trabalho, outras variáveis que dizem respeito às características sócio demográficas da população e às condições de trabalho. Essas variáveis estão elencadas no quadro 1:

Quadro 1 – Descrição de variáveis escolhidas da PNAD Contínua

Variável	**Descrição**
UF	Unidade da Federação
V1028	Peso trimestral com correção de não entrevista com pós estratificação pela projeção da população
V2007	Sexo
V2009	Idade do morador na data de referência
V2010	Cor ou raça
VD4001	Condição em relação à força de trabalho na semana de referência para pessoas de 14 anos ou mais de idade

VD4002	Condição de ocupação na semana de referência para pessoas de 14 anos ou mais de idade
VD4005	Pessoas desalentadas na semana de referência
VD4008	Posição na ocupação no trabalho principal
VD4019	Rendimento mensal habitual de todos os trabalhos para pessoas de 14 anos ou mais de idade (apenas para pessoas que receberam em dinheiro, produtos ou mercadorias em qualquer trabalho
VD4020	Rendimento mensal efetivo de todos os trabalhos para pessoas de 14 anos ou mais de idade (apenas para pessoas que receberam em dinheiro, produtos ou mercadorias em qualquer trabalho
VD 4035	Horas efetivamente trabalhadas na semana de referência em todos os trabalhos para pessoas de 14 anos ou mais de idade

Fonte: IBGE, Documentação da PNAD Contínua trimestral, Dicionário de variáveis (2020)

A tabela 1 mostra a composição da amostra cujo tamanho (n) é de 461.795 e 478.091 pessoas para 2021 e 2022, respectivamente. A proporção de mulheres é de 51,12% para o primeiro ano e de 51,13% para o segundo ano.

Tabela 1 – Composição do banco de dados por sexo

Sexo	**Número de Pessoas na amostra em 2021**	**Proporção de Pessoas na população em 2021 (%)**	**Número de Pessoas na amostra em 2022**	**Proporção de Pessoas na população em 2022 (%)**
Homens	223.980	48,88	232.029	48,17
Mulheres	237.815	51,12	246.062	51,13
TOTAL	461.795	100,0	478.091	100,0

Fonte: IBGE (2023) CV= 0,01

A tabela 2 mostra a composição da população em relação à raça com a proporção de cada grupo para os anos de 2021 e 2022.

Tabela 2 – Composição da população por raça

Cor ou raça	Número de Pessoas na população em 2021	Proporção de Pessoas na população em 2021 (%)	Número de Pessoas na população em 2022	Proporção de Pessoas na população em 2022 (%)
Branca	91.769.971	43,05	92.032.924	42,87
Preta	20.857.244	9,78	22.346.664	10,41
Amarela	1.583.125	0,74	1.523.512	0,71
Parda	98.061.660	46,00	97.832.973	45,57
Indígena	852.622	0,40	904.907	0,42
Ignorado	63.695	0,03	39.239	0,02
TOTAL	213.188.317	100,0	214.680.219	100,0

Fonte: IBGE (2023) CV = 0,01

Para um total de 107.758.094 pessoas, pode-se afirmar que 88,85% estão ocupadas, ou seja, exercem alguma atividade remunerada em 2021. Em 2022, esse percentual foi de 92,04% para uma população de 107.941.606 pessoas, mostrados na tabela 3.

Tabela 3 – Composição da população em relação à condição de ocupação

Condição de ocupação	Número de Pessoas na população em 2021	Proporção de Pessoas na população em 2021 (%)	Número de Pessoas na população em 2022	Proporção de Pessoas na população em 2022 (%)
Ocupados	95.747.458	88,85	99.369.771	92,06
Desocupados	12.010.636	11,15	8.571.835	7,94
TOTAL	107.758.094	100,0	107.941.606	100,0

Fonte: IBGE (2023) CV = 0,01

Pode-se dizer que a estimativa do rendimento mensal do trabalho principal, para o ano de 2021, foi de R$ 2493,70 e para 2022, esse valor foi de R$ 2902,60, em média.

A tabela 4 sintetiza os valores, em média, em relação ao sexo e a raça ou cor. Observa-se que, em média, o rendimento mensal do trabalho principal das mulheres é menor do que o dos homens. E as pessoas de cor preta, tem também os menores salários para os dois anos do estudo, seguido das pessoas de cor parda.

Tabela 4 – Renda média em relação ao sexo e à raça ou cor

Sexo **Raça ou cor**	**Valores médios em 2021 (R$)**	**Valores médios em 2022 (R$)**
Homem	2715,00	3194,90
Mulher	2188,70	2508,60
Branca	3169,90	3724,00
Preta	1913,00	2189,30
Amarela	3728,60	4383,90
Parda	1919,60	2226,10
Indígena	1782,20	2332,20
Ignorado	3388,40	6722,20

Fonte: IBGE (2023)

4.4 RESULTADOS

As análises foram feitas por meio de pacotes como *tidyverse,* desenvolvido por Wickham et al. (2019), e *survey*, desenvolvido por Lumley (2019) na interface R Studio, do software livre R. O pacote *survey* contém um conjunto de funções para o cálculo de estimativas de distintas estatísticas descritivas, aplicáveis a pesquisas com planos amostrais complexos.

Considerando a abordagem separada, foram calculados o tamanho amostral, o tamanho populacional estimado, a média amostral, o erro padrão da média amostral, a variância da média amostral, o coeficiente de variação da média amostral e o meff da variável VD4020: “rendimento efetivo no trabalho principal (R$)” a nível Brasil da PNAD Contínua do quarto trimestre dos anos de 2021 e 2022 considerando o efeito amostral. Os resultados estão descritos na tabela 5:

Tabela 5 – Resultados abordagem separada

	PNADC 2021	**PNADC 2022**
Tamanho amostral (n)	461.795	478.091
Tamanho populacional ($\widehat{N}$)	213.188.317	214.680.219
Média amostral ($\hat{\theta}$)	2493,70	2902,60
Erro padrão da média amostral (ep ($\hat{\theta}$))	22,63	28,77
Variância da média amostral (var ($\hat{\theta}$))	512,30	827,77
Coeficiente de Variação da média amostral (cv ($\hat{\theta}$))	0,01	0,01
Meff	19,36	22,36

Fonte: do próprio autor (2023)

Primeiramente, usou-se os métodos de ponderação da média para estimativas de pesquisas combinadas proposto por Fox (2010), o inverso do tamanho amostral e o inverso da variância, com o objetivo de dar uma estrutura para a meta análise. Maheswaran et al (2015) também usaram o inverso da variância. Os resultados obtidos estão registrados no quadro 5. Para os cálculos, usou-se as equações

$$\hat{\theta}_c = \sum_{d=1}^{D}\left(\frac{\frac{1}{n}*\hat{\theta}_d}{\sum_{d=1}^{D}\frac{1}{n}}\right) \qquad \text{e} \qquad \hat{\theta}_c = \sum_{d=1}^{D}\left(\frac{\frac{1}{var(\hat{\theta}_d)}*\hat{\theta}_d}{\sum_{d=1}^{D}\frac{1}{var(\hat{\theta}_d)}}\right).$$

Para a abordagem combinada, Dzikiti, Vieira e Girdler-Brown (2023) propuseram calcular a média combinada usando o inverso do cv e o inverso do meff como ponderação através das equações

$$\hat{\theta}_c = \sum_{d=1}^{D}\left(\frac{\frac{1}{cv(\hat{\theta}_d)}*\hat{\theta}_d}{\sum_{d=1}^{D}\frac{1}{cv(\hat{\theta}_d)}}\right) \qquad \text{e} \qquad \hat{\theta}_c = \sum_{d=1}^{D}\left(\frac{\frac{1}{\widehat{meff}(\hat{\theta}_d)}*\hat{\theta}_d}{\sum_{d=1}^{D}\frac{1}{\widehat{meff}(\hat{\theta}_d)}}\right).$$

Dzikiti (2019) propôs como uma alternativa simplificada para o cálculo da variância do estimador combinado a equação

$$\text{Var}\,(\hat{\theta}_c) = \frac{1}{D^2}\sum_{d=1}^{D}\text{Var}(\hat{\theta}_d)\,.$$

O método considerado para a estimação da variância do estimador combinado é simplificado e não leva em consideração as diferentes abordagens de ponderação.

Os resultados dos cálculos da média combinada usando os inversos do tamanho amostral, da variância, do cv e do meff usados como ponderação bem como os resultados do erro padrão, da variância e do cv da média combinada foram plotados na tabela 6.

Tabela 6 – Resultados abordagem combinada proposta

	Estimador combinado pelo inverso do tamanho da amostra	**Estimador combinado pelo inverso da variância**	**Estimador combinado pelo inverso do cv**	**Estimador combinado pelo inverso do meff**
Média combinada	2694,61	2650,02	2689,15	2712,88
Erro padrão média combinada	18,30	18,30	18,30	18,30
Variância média combinada	335,02	335,02	335,02	335,02
CV média combinada	0,01	0.01	0,01	0,01

Fonte: do próprio autor (2023)

5. CONSIDERAÇÕES FINAIS

Os resultados obtidos por Dzikiti, Vieira e Girdler-Brown (2023) quando, inicialmente, usando de simulações, compararam as propriedades de convergência de estimativas combinadas para três diferentes estratégias de ponderação quando usam o inverso do tamanho da amostra e o inverso da variância, concordaram com os resultados obtidos por Fox (2010), ou seja, a estimativa da média combinada convergiu para a média da população finita. Este é um resultado ideal porque na metanálise, o uso do inverso da variância para estudos de peso resulta na variância mínima para a estimativa do tamanho do efeito combinado.

Em seguida, ao investigarem o uso de diferentes estratégias de ponderação para combinar pesquisas sob diferentes modelos de amostragem, no uso da amostragem aleatória simples, da amostragem estratificada e da amostragem por conglomerados, através de simulação e com diferentes tamanhos de população finita, Dzikiti, Vieira e Girdler-Brown (2023) concluíram que, sob amostragem estratificada e amostragem por conglomerado, ponderações consistentes como o inverso do cv e o inverso do meff,

resultaram em valores de EQM diferentes das estratégias de ponderação mais simples. Resultados de outros ponderadores para a abordagem combinada pode ser consultado em Dzikiti, Vieira e Girdler-Brown (2023).

O resultado do estimador combinado tem como parâmetro alvo uma população fictícia e dinâmica dos anos de 2021 e 2022. E, ao aplicar estas estratégias de ponderação em dados da PNAD Contínua de 2021 e 2022, obteve-se um erro padrão do estimador de 18,30. Para a PNAD Contínua de 2021, essa medida foi de 22,63 enquanto que para a PNAD 2022, esse valor foi de 28,77.

Conclui-se, finalmente, que combinar as pesquisas resultou em estimativas melhores com erros padrão menores do que seriam observados quando somente uma pesquisa é utilizada, confirmando os resultados obtidos por Dzikiti, Vieira e Girdler-Brown (2023).

6. REFERENCIAS BIBLIOGRÁFICAS

Bolfarine, H; Bussab, W. O. **Elementos de Amostragem**: São Paulo, Edgard'Blucher, 2005.

Braga D, Assuncao G. **PNADcIBGE: Downloading, Reading and Analyzing PNADC Microdata_**. R package version 0.7.2, 2023. Disponível em: https://CRAN.R-project.org/package=PNADcIBGE.

Cochran, W. G. **Sampling techniques**. 2. ed., New York, John Wiley & Sons, Inc., 1995.

Deeks J.J., Higgins J.P.T, Altman D.G. **Análise de dados e realização de meta-análises**. Manual Cochrane para revisões sistemáticas de intervenções versão 6.3, Cochrane, 2022. Disponível em www.training.cochrane.org/handbook.

Dzikiti, L. N. et al, "Comparing **approaches for combining data collected from multiple complex surveys, adjusting for clustering and stratification".** Tese de Doutorado. University of Pretoria, 2019.

Fox K. **A Framework for the Meta-Analysis of Survey Data**. In: JSM Proceedings, Survey Research Methods Section. 2010.

Kish L. **Survey Sampling. Classics**. New York: Wiley; 1995.

Lumley, T. **survey: analysis of complex survey samples**. R package version 4.2, 2019.

IBGE - Pesquisa Nacional por Amostra de Domicílos Contínua- – Notas técnicas. 2018.

IBGE - Pesquisa Nacional por Amostra de Domicílos Contínua- – Dicionário de variáveis. 2022.

QUINTSLR, M. M. M. et al. **Sistema Integrado de Pesquisas Domiciliares (SIPD).** Rio de Janeiro: Coordenação de Trabalho e Rendimento. Diretoria de Pesquisas, Instituto Brasileiro de Geografia e Estatística, 2007. (Texto para Discussão n. 24)

Roberts G, Binder D. **Analyses Based on Combining Similar Information from Multiple Surveys.** In: JSM Proceedings, Survey Research Methods Section.2009.

Schenker N, Gentleman JF, Rose D, Hing E, Shimizu IM. **Combining estimates from complementary surveys: a case study using prevalence estimates from national health surveys of households and nursing homes**. PublicHealth Reports, 2002.

Schenker, N. and T.E. Raghunathan. **Combining Information from Multiple Surveys to Enhance Estimation of Measures of Health**, Statistics in Medicine, 2007.

Skinner, C.J., Holt, T. e Smith, T.M.F. **Analysis of Complex Surveys**. Chichester: Wiley, 1989.

Trovão, C. J. B. M. **Por dentro da PNAD contínua [recurso eletrônico]: uma introdução ao tratamento de dados usando R** – Dados eletrônicos (1 arquivo: 78 KB). Natal, RN: EDUFRN, 2022.

Wickham H, Vaughan D, Girlich M. **_tidyr: Tidy Messy Data_. R package version 1.3.**0, 2023. Disponível em https://CRAN.R-project.org/package=tidyr .

7. ANEXOS

7.3 CÓDIGO EM R DA PNAD CONTÍNUA 2021 COM SAÍDAS

```
# variáveis
vars<-c("Ano","Trimestre","UF","UPA","Estrato","V1028","V2007","V2009",
        "V2010","VD4001","VD4002","VD4005","VD4008","VD4019","VD4020",
        "VD4035")
# dados brutos
dados<-get_pnadc(year = 2021, quarter = 4,design = FALSE,
                 labels = TRUE, vars = vars)
# número de homens e mulheres na população
dados %>% count(V2007, wt = V1028)

##   V2007          n
##   <fct>      <dbl>
## 1 Homem  104201183.
## 2 Mulher 108987134.

# número de homens e mulheres na amostra
dados %>% group_by(V2007) %>% summarise(n = n())

##   V2007      n
##   <fct>  <int>
## 1 Homem  223980
## 2 Mulher 237815

# dados com desenho amostral
dados_dsg <-get_pnadc(2021,4,vars = vars)
class(dados_dsg)

# Formato data.frame / tibble
dados_df <- as_tibble(dados_dsg$variables)
class(dados_df)

# sexo (V2007)

svytotal(~V2007,  dados_dsg,  na.rm = T)

##                            total     SE
## V2007Homem  104201183 0.0422
## V2007Mulher 108987134 0.0353


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

```
##                          mean SE
## V2007Homem  0.48878  0
## V2007Mulher 0.51122  0

cv(object = prop_sexo)

##   V2007Homem  V2007Mulher
## 4.613995e-11 4.411392e-11

# raça ou cor
svytotal(~V2010, dados_dsg, na.rm = T)

##                       total             SE
## V2010Branca   91769971 441237
## V2010Preta    20857244 288169
## V2010Amarela   1583125  87004
## V2010Parda    98061660 414972
## V2010Indígena   852622  42921
## V2010Ignorado    63695  14764


(prop_race<-svymean(~V2010, dados_dsg, na.rm = T))


##                            mean     SE
## V2010Branca   0.43046435 0.0021
## V2010Preta    0.09783484 0.0014
## V2010Amarela  0.00742595 0.0004
## V2010Parda    0.45997671 0.0019
## V2010Indígena 0.00399939 0.0002
## V2010Ignorado 0.00029877 0.0001

cv(object = prop_race)

##   V2010Branca    V2010Preta  V2010Amarela    V2010Parda V2010Indígena
##   0.004808076   0.013816251   0.054957398   0.004231743   0.050339613
## V2010Ignorado
##   0.231791007

# ocupação (VD4002)
svytotal(~VD4002, dados_dsg, na.rm = T)

##                          total     SE
## VD4002Pessoas ocupadas    95747458 218221
## VD4002Pessoas desocupadas 12010636 128933

# proporção
(prop_ocup<-svymean(~VD4002, dados_dsg, na.rm=TRUE) )

##                          mean     SE
## VD4002Pessoas ocupadas    0.88854 0.0012
## VD4002Pessoas desocupadas 0.11146 0.0012

cv(object = prop_ocup)
```

```
##    VD4002Pessoas ocupadas VD4002Pessoas desocupadas
##              0.00132243               0.01054227

# estimativa rendimento mensal efetivo do trabalho principal em relação a raça ou cor
svymean(~VD4020, subset(dados_dsg, V2010 == "Preta" ), na.rm = TRUE)

##             mean     SE
## VD4020 1913 27.616

svymean(~VD4020, subset(dados_dsg, V2010 == "Branca" ), na.rm = TRUE)

##             mean     SE
## VD4020 3169.9 39.305

svymean(~VD4020, subset(dados_dsg, V2010 == "Amarela" ),na.rm = TRUE)

##              mean     SE
## VD4020 3728.6 215.37

svymean(~VD4020, subset(dados_dsg, V2010 == "Parda" ), na.rm = TRUE)

##             mean   SE
## VD4020 1919.6 15.5

svymean(~VD4020, subset(dados_dsg, V2010 == "Indígena" ),na.rm = TRUE)

##              mean     SE
## VD4020 1782.2 131.93

svymean(~VD4020, subset(dados_dsg, V2010 == "Ignorado" ),na.rm = TRUE)

##              mean     SE
## VD4020 3388.4 987.82

# estimativa rendimento efetivo mensal do trabalho principal em relação ao sexo

 svymean(~VD4020, subset(dados_dsg, V2007 == "Homem" ), na.rm = TRUE)

##             mean     SE
## VD4020 2715 26.753

svymean(~VD4020, subset(dados_dsg, V2007 == "Mulher" ), na.rm = TRUE)

##              mean     SE
## VD4020 2188.7 22.665

# renda
(totalrenda <- svytotal(~VD4020, dados_dsg, na.rm = T))


##                   total         SE
## VD4020 2.3374e+11 2245496973

 cv(object = totalrenda)
```

```
##      VD4020
## 0.009606829

confint(object = totalrenda)

##                          2.5 %       97.5 %
## VD4020 229338561671 238140748061

# media
(mediarenda <- svymean(~VD4020, dados_dsg, na.rm = T))


##               mean     SE
## VD4020 2493.7 22.634

cv(object = mediarenda)

##      VD4020
## 0.009076269

confint(object = mediarenda)

##                    2.5 %   97.5 %
## VD4020 2449.339 2538.061

# calculo do meff
# dados brutos=> sem desenho amostral para
dadosbrutos <-  get_pnadc(2021, quarter = 4,
                    design = FALSE, labels = T,
                    vars = c("Ano","Trimestre","UF",
                    "UPA","Estrato","V1028", "VD4020"))

renda <- dadosbrutos %>%  pull(VD4020)

 var(renda, na.rm = TRUE)

## [1] 12222380

 mean(renda, na.rm = TRUE)

## [1] 2246.35
```

7.4 CÓDIGO EM R DA PNAD CONTÍNUA 2022 COM SAÍDAS

```
# variáveis
vars<-c("Ano","Trimestre","UF","UPA","Estrato","V1028","V2007","V2009",
       "V2010","VD4001","VD4002","VD4005","VD4008","VD4019","VD4020",
        "VD4035")

# dados brutos
dados<-get_pnadc(year = 2022, quarter = 4,design = FALSE,
              labels = TRUE, vars = vars)

# numero de homens e mulheres na população
dados %>% count(V2007, wt = V1028)

##   V2007          n
##   <fct>      <dbl>
## 1 Homem  104910367.
## 2 Mulher 109769852.

# numero de homens e mulheres na amostra
dados %>% group_by(V2007) %>% summarise(n = n())
##   V2007      n
##   <fct>  <int>
## 1 Homem  232029
## 2 Mulher 246062

# dados com desenho amostral
dados_dsg <-get_pnadc(2022,4,vars = vars)


# Formato data.frame / tibble
dados_df <- as_tibble(dados_dsg$variables)


# sexo (V2007)
svytotal(~V2007, dados_dsg, na.rm = T)

##                 total     SE
## V2007Homem  104910367 0.0517
## V2007Mulher 109769852 0.0409

#proporção
(prop_sexo<-svymean(~V2007, dados_dsg, na.rm = T))
##                mean SE
```

```
## V2007Homem  0.48868  0
## V2007Mulher 0.51132  0

cv(object = prop_sexo)

##   V2007Homem  V2007Mulher
## 7.585087e-11 7.249297e-11

# raça ou cor (V2010)
svytotal(~V2010, dados_dsg, na.rm = T)

##                 total     SE
## V2010Branca   92032924 446685
## V2010Preta    22346664 227471
## V2010Amarela   1523512  63877
## V2010Parda    97832973 393707
## V2010Indígena   904907  43829
## V2010Ignorado    39239  10828

# proporçao
(prop_race<-svymean(~V2010, dados_dsg, na.rm = T))
##                   mean     SE
## V2010Branca   0.42869774 0.0021
## V2010Preta    0.10409280 0.0011
## V2010Amarela  0.00709666 0.0003
## V2010Parda    0.45571489 0.0018
## V2010Indígena 0.00421514 0.0002
## V2010Ignorado 0.00018278 0.0001

cv(object = prop_race)

##   V2010Branca    V2010Preta  V2010Amarela    V2010Parda V2010Indígena
##   0.004853536   0.010179204   0.041927332   0.004024282   0.048435124
## V2010Ignorado
##   0.275961655

# ocupação
svytotal(~VD4002, dados_dsg, na.rm = T)

##                            total     SE
## VD4002Pessoas ocupadas    99369771 221418
## VD4002Pessoas desocupadas  8571835 103989

# proporção
(prop_ocup<-svymean(~VD4002, dados_dsg, na.rm=TRUE))
```

```
##                             mean    SE
## VD4002Pessoas ocupadas    0.920588 0.001
## VD4002Pessoas desocupadas 0.079412 0.001
```

```
cv(object = prop_ocup)
```

```
##    VD4002Pessoas ocupadas VD4002Pessoas desocupadas
##              0.001034914              0.011997330
```

```
# estimativa rendimento mensal efetivo do trabalho principal em relação à raça
svymean(~VD4020, subset(dados_dsg,  V2010 == "Preta" ),  na.rm = TRUE)
```

```
##          mean     SE
## VD4020 2189.3 27.804
```

```
svymean(~VD4020, subset(dados_dsg, V2010 == "Branca" ), na.rm = TRUE)
```

```
##        mean     SE
## VD4020 3724 51.169
```

```
svymean(~VD4020, subset(dados_dsg, V2010 == "Amarela" ), na.rm = TRUE)
```

```
##          mean     SE
## VD4020 4383.9 295.85
```

```
svymean(~VD4020, subset(dados_dsg, V2010 == "Parda" ), na.rm = TRUE)
```

```
##          mean     SE
## VD4020 2226.1 17.855
```

```
svymean(~VD4020, subset(dados_dsg, V2010 == "Indígena" ), na.rm = TRUE)
```

```
##          mean     SE
## VD4020 2332.2 158.04
```

```
svymean(~VD4020, subset(dados_dsg, V2010 == "Ignorado" ), na.rm = TRUE)
```

```
##          mean     SE
## VD4020 6722.2 2067.2
```

```
#estimativa rendimento mensal efetivo do trabalho principal em relação ao sexo
svymean(~VD4020, subset(dados_dsg, V2007 == "Homem" ), na.rm = TRUE)
```

```
##          mean     SE
## VD4020 3194.9 37.924
```

```
svymean(~VD4020, subset(dados_dsg, V2007 == "Mulher" ),na.rm = TRUE)
```

```
##          mean     SE
## VD4020 2508.6 21.709
```

```
# renda
 (totalrenda <- svytotal(~VD4020, dados_dsg, na.rm = T))


##            total        SE
## VD4020 2.8355e+11 2943576346

cv(object = totalrenda)

##     VD4020
## 0.01038118

confint(object = totalrenda)

##            2.5 %       97.5 %
## VD4020 2.7778e+11 289318608912


(mediarenda <- svymean(~VD4020, dados_dsg, na.rm = T))
##          mean     SE
## VD4020 2902.6 28.771

cv(object = mediarenda)

##     VD4020
## 0.00991218

confint(object = mediarenda)

##          2.5 %   97.5 %
## VD4020 2846.247 2959.029

# dados brutos sem desenho amostral para cálculo do meff

dadosbrutos <-  get_pnadc(2022, quarter = 4,
          design = FALSE, labels = T,
          vars = c("Ano","Trimestre","UF",
                "UPA","Estrato","V1028",
                "VD4020"))

renda <- dadosbrutos %>%  pull(VD4020)
var(renda, na.rm = TRUE)

## [1] 17698038

mean(renda, na.rm = TRUE)

## [1] 2642.83
```