

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

DAIANE LIMA DO NASCIMENTO

AUTOAVALIAÇÃO DA SAÚDE DA POPULAÇÃO BRASILEIRA RESIDENTE NO
NORDESTE E FATORES SOCIODEMOGRÁFICOS: UM ESTUDO DE
ASSOCIAÇÃO COM BASE NA PESQUISA NACIONAL DE SAÚDE 2013

ORIENTADOR: Prof. Dr. Damião Nóbrega da Silva

NATAL/RN
2018

DAIANE LIMA DO NASCIMENTO

AUTOAVALIAÇÃO DA SAÚDE DA POPULAÇÃO BRASILEIRA RESIDENTE NO
NORDESTE E FATORES SOCIODEMOGRÁFICOS: UM ESTUDO DE
ASSOCIAÇÃO COM BASE NA PESQUISA NACIONAL DE SAÚDE 2013

Monografia apresentada ao Departamento
de Estatística da Universidade Federal do
Rio Grande do Norte, em cumprimento
com as exigências legais para obtenção do
título de formação em Estatística.

ORIENTADOR: Prof. Dr. Damião Nóbrega da Silva

NATAL/RN
2018

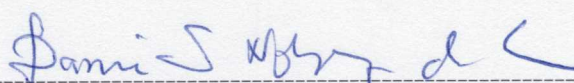
DAIANE LIMA DO NASCIMENTO

AUTOAVALIAÇÃO DA SAÚDE DA POPULAÇÃO BRASILEIRA RESIDENTE NO
NORDESTE E FATORES SOCIODEMOGRÁFICOS: UM ESTUDO DE
ASSOCIAÇÃO COM BASE NA PESQUISA NACIONAL DE SAÚDE 2013

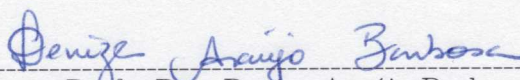
Monografia apresentada ao Departamento
de Estatística da Universidade Federal do
Rio Grande do Norte, em cumprimento
com as exigências legais para obtenção do
título de formação em Estatística.

MONOGRAFIA APROVADA EM 17/12/2018

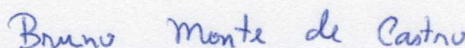
BANCA EXAMINADORA



Prof. Dr. Damiano Nóbrega da Silva
(Orientador)



Profa. Dra. Denize Araújo Barbosa
(Membro da Banca)



Profa. Dr. Bruno Monte De Castro
(Membro da Banca)

NATAL/RN
2018

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Ronaldo Xavier de Arruda - CCET

Nascimento, Daiane Lima do.

Autoavaliação da saúde da população brasileira residente no Nordeste e fatores sociodemográficos: um estudo de associação com base na Pesquisa Nacional de Saúde 2013 / Daiane Lima do Nascimento. - 2018.

76f.: il.

Monografia (Bacharelado em Estatística) - Universidade Federal do Rio Grande do Norte, Centro de Ciências Exatas e da Terra, Departamento de Estatística. Natal, 2018.

Orientador: Damião Nóbrega da Silva.

1. Estatística - Monografia. 2. Amostragem complexa - Monografia. 3. Teste de independência - Monografia. 4. Correção de Rao-Scott - Monografia. 5. Não-resposta - Monografia. 6. Calibração - Monografia. I. Silva, Damião Nóbrega da. II. Título.

RN/UF/CCET

CDU 519.2

Agradecimentos

A Deus, por cuidar de cada detalhe em minha vida. A todos os professores do departamento de Estatística desta universidade, em especial: ao professor Damião Nóbrega da Silva, pela paciência, confiança e orientações que proporcionaram a realização deste trabalho; aos professores Pledson Guedes de Medeiros e André Luis Santos de Pinho pela atenção demonstrada ao longo do curso; aos professores Bruno Monte de Castro e Denize Araújo Barbosa que aceitaram participar desta banca. A minha família, amigos e namorado, pelo apoio e incentivo nos dias difíceis. Obrigada a todos. Vocês tornaram esta jornada significativa.

“Não to mandei eu? Esforça-te e tem bom ânimo; não pases, nem te espantes, porque o SENHOR, teu Deus, é contigo, por onde quer que andares.”
(Bíblia Sagrada, Josué 1,9)

Resumo

A autoavaliação de saúde é um indicador subjetivo, mas está associado a estilos de vida, comorbidades e mortalidade. Uma fonte de dados que possibilita a estimação desse indicador para a população brasileira é a Pesquisa Nacional de Saúde, 2013, que se baseia em uma amostra estratificada com seleção de conglomerados em múltiplos estágios. Usando dados desta pesquisa, este trabalho busca testar a hipótese de independência entre o estado de saúde auto-referido e variáveis sociodemográficas na população adulta do nordeste brasileiro, além de descrever essas associações. A subamostra analisada abrange cerca de 19.419 domicílios permanentes particulares e 18.305 respondentes. Aplicando-se testes de independência incorporando-se a estrutura do plano amostral foram detectadas fortes evidências estatísticas de associação da saúde auto-referida com as variáveis sexo, faixa etária, nível de instrução e situação habitacional e de associação moderada entre o indicador de saúde e raça. Em geral, observa-se que moradores adultos do sexo feminino, que tem mais anos de idade, não possuem instrução ou têm apenas o fundamental incompleto e residem em região rural estão associados à autoavaliação de saúde regular ou ruim/muito ruim. Também, moradores do sexo masculino, com maior escolaridade e que vivem em regiões urbanas estão associados à saúde auto-referida como boa ou muito boa.

Palavras-chave: Amostragem complexa. Teste de independência. Correção de Rao-Scott. Não-resposta. Calibração.

Abstract

Health self-assessment is a subjective indicator, but it is associated with lifestyles, comorbidities, and mortality. One source of data that makes it possible to estimate this indicator for the Brazilian population is the National Health Survey 2013, which is based on a stratified multi-stage cluster sample. Using data from this survey, this work seeks to test the hypothesis of independence between the self-reported health status and sociodemographic variables in the adult population of the Brazilian Northeast, in addition to describing these associations. The subsample analyzed covers approximately 19,419 private permanent households and 18,305 respondents. Applying independence tests incorporating the sampling plan structure, strong statistical evidence of self-reported health association was detected with the variables gender, age group, educational level and housing status and a moderate association between the health indicator and race. In general, it is observed that adult female residents, who are older, have no or lower education and reside in rural areas are associated with regular or poor or very poor health self-assessment. Also, male residents with higher education living in urban areas are associated with good or very good self-reported health.

Keywords: Complex sampling. Independence test. Rao-Scott correction. Nonresponse. Calibration.

Lista de ilustrações

Figura 1 – Diagramas de caixa dos pesos amostrais dos moradores com correções para não respostas e calibração por Unidade Federativa da Região Nordeste, 2013.	36
Figura 2 – Histogramas das idades dos moradores da região Nordeste - PNS 2013. Em (a), as densidades são estimadas usando as proporções amostrais de indivíduos (sem correção) em cada classe etária e, em (b), as densidades são estimadas com as estimativas das proporções ponderadas pelos pesos amostrais, Região Norteste, 2013.	38
Figura 3 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde condicional ao sexo, Região Nordeste, 2013.	41
Figura 4 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde condicional à faixa etária, Região Nordeste, 2013.	44
Figura 5 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde e raça, Região Nordeste, 2013.	47
Figura 6 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde e nível de instrução, Região Nordeste, 2013.	50
Figura 7 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde e situação habitacional, Região Nordeste, 2013.	53

Lista de Tabelas

Tabela 1 – Tabela de contingência populacional das variáveis x e y	29
Tabela 2 – Tabela de contingência estimada ponderada das variáveis x e y	30
Tabela 3 – Estatísticas descritivas para os pesos amostrais dos moradores com correções para não resposta e calibração, Região Nordeste, 2013.	35
Tabela 4 – Número de entrevistados (n), proporção amostral, proporção ponderada, coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias das variáveis da análise, Região Nordeste, 2013.	37
Tabela 5 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo o sexo e a autoavaliação da saúde, Região Nordeste, 2013.	40
Tabela 6 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional ao sexo, Região Nordeste, 2013.	40
Tabela 7 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e o sexo na população alvo, Região Nordeste, 2013.	40
Tabela 8 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo a faixa etária e a autoavaliação da saúde, Região Nordeste, 2013.	42
Tabela 9 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional à faixa etária, Região Nordeste, 2013.	43
Tabela 10 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e a faixa etária na população alvo, Região Nordeste, 2013.	43
Tabela 11 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo a raça e a autoavaliação da saúde, Região Nordeste, 2013.	45

Tabela 12 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional à raça, Região Nordeste, 2013.	46
Tabela 13 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e raça na população alvo, Região Nordeste, 2013.	46
Tabela 14 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo o nível de instrução e a autoavaliação da saúde, Região Nordeste, 2013.	48
Tabela 15 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional ao nível de instrução, Região Nordeste, 2013.	49
Tabela 16 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e o nível de instrução na população alvo, Região Nordeste, 2013.	49
Tabela 17 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo a situação habitacional e a autoavaliação da saúde, Região Nordeste, 2013.	51
Tabela 18 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional à situação habitacional, Região Nordeste, 2013.	52
Tabela 19 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e situação habitacional na população alvo, Região Nordeste, 2013.	52

Sumário

1	INTRODUÇÃO	17
2	METODOLOGIA	19
2.1	Características dos dados do estudo	19
2.1.1	<i>Plano amostral</i>	19
2.1.2	<i>Método de coleta de dados</i>	20
2.1.3	<i>Estrutura dos microdados</i>	20
2.1.4	<i>Variáveis</i>	21
2.1.5	<i>Pesos amostrais</i>	21
2.2	Métodos de estimação	25
2.2.1	<i>Estimação de totais e médias</i>	25
2.2.2	<i>Estimação em domínios</i>	26
2.2.3	<i>Estimação de frequências e proporções</i>	27
2.2.4	<i>Estimação de variâncias</i>	28
2.3	Teste de independência	29
3	ANÁLISE DOS RESULTADOS	35
3.1	Análise das distribuições univariadas	35
3.2	Estudo de associações	39
3.2.1	<i>Associação entre autoavaliação de saúde e sexo</i>	39
3.2.2	<i>Associação entre autoavaliação de saúde e faixa etária</i>	42
3.2.3	<i>Associação entre a autoavaliação de saúde e raça</i>	45
3.2.4	<i>Associação entre autoavaliação de saúde e nível de instrução</i>	48
3.2.5	<i>Associação entre autoavaliação de saúde e situação habitacional</i>	51
4	CONSIDERAÇÕES FINAIS	55
	REFERÊNCIAS	57
	APÊNDICES	59
A	VARIÁVEIS	61
B	CÓDIGO R	62

1 Introdução

Um método de avaliação do estado de saúde de uma população humana consiste em entrevistar os indivíduos desta população acerca da percepção que eles têm sobre a própria saúde. Estas medidas de percepção, também conhecidas como medidas *auto-referidas* ou de *auto-avaliação*, podem ser tomadas para formar um indicador global de saúde da população. Duas vantagens deste tipo de medida são a facilidade de obtenção em populações com elevado número de indivíduos e a possibilidade de comparação com inquéritos internacionais (THEME-FILHA; SZWARCOWALD; SOUZA-JÚNIOR, 2008; THEME-FILHA; SZWARCOWALD; SOUZA-JÚNIOR, 2005). Apesar deste indicador ser subjetivo, estudos apontam que o mesmo está associado com comorbidades e estilos de vida além de ser preditor de mortalidade (BARROS et al., 2009; IDLER; BENYAMINI, 1997; BURSTRÖM; FREDLUND, 2001).

Uma fonte rica de informação sobre a saúde auto-referida da população brasileira adulta é a Pesquisa Nacional de Saúde (PNS). Esta pesquisa é um estudo transversal que foi criado pelo Ministério da Saúde, a Fundação Osvaldo Cruz e o Instituto Brasileiro de Geografia e Estatística (IBGE), para ter periodicidade de cinco anos a partir da primeira edição, que foi realizada em 2013. De acordo com (SZWARCOWALD et al., 2014),

a PNS tem como objetivo geral o de suprir dados em âmbito nacional que permitam caracterizar a situação de saúde e os estilos de vida da população brasileira, bem como a atenção à saúde, no que se refere ao acesso e uso dos serviços de saúde, às ações preventivas, à continuidade dos cuidados, e ao financiamento da assistência de saúde.

A população alvo da pesquisa é o conjunto dos moradores de 18 anos ou mais de idade que residem em domicílios particulares permanentes do Brasil, excluindo-se aqueles moradores residindo em domicílios especiais, tais como quartéis, bases militares, alojamentos, acampamentos, embarcações, penitenciárias, colônias penais, presídios, cadeias, asilos, orfanatos, conventos e hospitais.

A PNS se baseia em uma amostra probabilística de moradores da população, selecionada por amostragem estratificada com seleção de conglomerados em três estágios dentro de cada estrato (FREITAS, 2014; SOUZA-JÚNIOR et al., 2015). Estimativas de vários indicadores sobre percepção da saúde, estilos de vida e doenças crônicas da PNS 2013 são apresentadas em IBGE (2014). As estimativas fornecidas referem-se não apenas ao território nacional, mas também às regiões do país, unidades da federação e outras subpopulações de interesse, tais como grupos de sexo, faixa etária, raça, escolaridade e situação do domicílio. Os indicadores sobre a percepção da saúde são prevalências de moradores com 18 anos ou mais com auto-avaliação da saúde boa ou muito boa, que têm sintomas de angina e que usam algum recurso para locomoção. Os indicadores sobre estilos de vida

abordam consumo alimentar, uso de álcool, atividade física, hábito de assistência à televisão e tabagismo. Por último, os indicadores das doenças crônicas referem-se a doenças tais como hipertensão arterial, diabetes, colesterol, asma, doenças cardiovasculares, acidente vascular cerebral, depressão, câncer, dentre outras doenças.

Dentre os resultados apresentados em IBGE (2014) sobre a auto-avaliação da saúde, observa-se que as estimativas das prevalências de moradores com 18 ou mais com auto-avaliação boa ou muito boa variam consideravelmente entre diferentes subpopulações de moradores, notadamente aqueles grupos definidos pelas categorias de variáveis sócio-demográficas. Neste trabalho, busca-se entender melhor estas diferenças através dos seguintes objetivos principais:

- i. testar as hipóteses de que a auto-avaliação da saúde independe de fatores sócio-demográficos na população residente adulta da região Nordeste;
- ii. avaliar como se dão as associações nos casos em que sejam obtidas evidências estatísticas para rejeição das hipóteses de independência em i.

Os testes das hipóteses de independência e estudo das associações nos objetivos i. e ii. serão realizados aplicando-se métodos estatísticos que incorporem correções para aspectos inerentes ao plano amostral da PNS 2013. Os fatores sócio-demográficos de interesse neste estudo são: sexo, idade, raça e escolaridade dos moradores, bem como a situação habitacional dos domicílios em que os moradores residem.

A estrutura deste trabalho é a seguinte: No Capítulo 2, são fornecidos maiores detalhes sobre o plano amostral da PNS 2013, os microdados usados desta pesquisa para o trabalho, o processo de cálculo dos pesos amostrais da PNS e uma revisão de métodos de estimação e testes de independência que são apropriados para dados amostrais complexos. No Capítulo 3, são apresentados as análises realizadas no trabalho. Finalmente, no Capítulo 4, são apresentadas algumas considerações finais.

2 Metodologia

2.1 Características dos dados do estudo

2.1.1 Plano amostral

A PNS é uma amostra probabilística estratificada com seleção de conglomerados em múltiplos estágios. Ela faz parte do Sistema Integrado de Pesquisas Domiciliares (SIPD) do IBGE, que se baseia em um Cadastro Mestre de setores censitários (áreas geográficas ou grupo de áreas geográficas menores). Esse cadastro possui abrangência nacional, excluindo-se apenas setores com número pequeno de domicílios ou considerados especiais (IBGE, 2014).

Dois características do SIPD que são importantes para um melhor entendimento do plano amostral da PNS são:

1. estratificação: os setores censitários do Cadastro Mestre são estratificados (SOUZA-JÚNIOR et al., 2015) por:
 - unidades da federação, com subdivisão adicional de cada unidade em: capital região metropolitana ou Região Integrada de Desenvolvimento Econômico (em alguns estados) e demais municípios;
 - regiões geográficas, dadas pela subdivisão das capitais e grandes cidades em regiões menores;
 - situações dos setores, em que as regiões geográficas são subdivididas em áreas urbanas ou rurais;
 - grupos de maior homogeneidade, em que os estratos anteriores são subdivididos para formar grupos homogêneos de acordo, por exemplo, com a renda média domiciliar.
2. Amostra Mestra: essa amostra é uma amostra estratificada do Cadastro Mestre, em que setores censitários são selecionados em cada estrato por amostragem aleatória com probabilidades proporcionais ao número de domicílios permanentes privados dos setores censitários.

A PNS mantém a mesma estratificação da Amostra Mestra e, a partir dessa, adota o seguinte esquema de amostragem de conglomerados em três estágios em cada estrato:

- primeiro estágio: seleção de uma amostra de setores censitários (unidades de amostragem primárias - UAP) da Amostra Mestra por amostragem aleatória simples;

- segundo estágio: dentro de cada setor censitário selecionado no primeiro estágio, seleção de domicílios particulares usando amostragem aleatória simples;
- terceiro estágio: dentro de cada domicílio selecionado no segundo estágio, seleção de um adulto de 18 anos ou mais com mesma probabilidade entre todos os adultos elegíveis do domicílio.

2.1.2 Método de coleta de dados

O método de coleta de dados da PNS consiste em entrevistas pessoais com aplicação de um questionário. Há, também, uma parte dos dados que são obtidas através de exames laboratoriais. A estrutura do questionário aplicado pela PNS pode ser subdividida em três módulos. No primeiro módulo, questões relativas a características do domicílio são respondidas, por exemplo, pelo chefe da unidade domiciliar. No segundo módulo, são abordadas questões sobre todos os residentes do domicílio. Finalmente, no terceiro módulo, características do residente selecionado para a amostra são investigadas.

2.1.3 Estrutura dos microdados

Os microdados da PNS 2013 são disponibilizados publicamente pelo IBGE. Um dos links pelos quais os dados podem ser obtidos é (<https://www.ibge.gov.br/estatisticas-novoportal/sociais/saude/>). Instruções e códigos sobre importações e análises dos dados da pesquisa usando o software R (R Core Team, 2018) foram preparadas por Anthony J. Damico e Dr. Djalma Galvão Carneiro Pessoa. Estas informações são disponibilizadas no site *Analyze Survey Data for Free*, que pode ser acessado pelo link (<http://asdfree.com/pesquisa-nacional-de-saude-pns.html>).

Para processar e realizar as análises deste trabalho, um código R que representa uma adaptação dos programas fornecidos no site *Analyze Survey Data for Free* é dado no Apêndice B. Após a importação dos microdados da PNS 2013, verifica-se um total de 6.062 unidades primárias de amostragem (UPAs) em 575 estratos para todo o país. As informações listadas referem-se a 64.308 domicílios e 60.202 adultos que foram selecionados na amostra e tiveram entrevista realizada.

Em virtude do interesse deste trabalho corresponder à região Nordeste, a subamostra resultante da PNS 2013 refere-se a 1.914 UPAs em 191 estratos, abrangendo cerca de 19.419 domicílios e um total de 18.305 respondentes da população alvo da região Nordeste. Estas quantidades de UPAs e domicílios nos dados obtidos apresentam pequenas diferenças das mesmas quantidades que são divulgadas em (SOUZA-JÚNIOR et al., 2015), em que os números totais de UPAs são 6.069 (Brasil) e 1.916 (Nordeste) e os número totais de domicílios são 64.348 (Brasil) e 19.431 (Nordeste).

2.1.4 Variáveis

As variáveis da PNS 2013 que foram julgadas necessárias para as análises a serem realizadas neste trabalho são listadas no Apêndice A. Nesta lista, pode-se obter também os respectivos códigos das variáveis que são usados no código R do Apêndice B, bem como os rótulos para as categorias das variáveis categorizadas. Algumas observações sobre uma parte destas variáveis são:

- Auto-percepção da saúde: medida obtida a partir da pergunta do questionário “Em geral, como o(a) sr(a) avalia a sua saúde?”. As opções de resposta para os entrevistados são: “muito boa”, “boa”, “regular”, “ruim” ou “muito ruim”. Esta variável foi reclassificada para três categorias, em que as categorias “muito boa” e “boa” foram classificadas como “boa” e as categorias “ruim” e “muito ruim” como “ruim”;
- Sexo do morador: com as categorias “Masculino” e “Feminino”;
- Idade do morador: foi obtida uma versão da variável idade do morador categorizada nas faixas etárias “18 a 29 anos”, “30 a 59 anos”, “60 a 64 anos”, “65 a 74 anos” e “75 anos ou mais”;
- Raça do morador: as categorias foram reclassificadas para “Branca”, “Preta”, “Parda” e “Amarela ou Indígena”;
- Nível de escolaridade do morador: as categorias foram reclassificadas para “Sem instrução ou fundamental incompleto”, “Fundamental completo ou médio incompleto”, “Médio completo ou superior incompleto” e “Superior completo”;
- Situação censitária do domicílio: com categorias “Urbano” e “Rural”.

2.1.5 Pesos amostrais

A PNS disponibiliza pesos amostrais para serem usados como fatores de ponderação ou de expansão da amostra. Estes fatores permitem análises dos dados da pesquisa incorporando-se os aspectos do plano amostral nas estimativas a serem produzidas. Dois conjuntos de pesos amostrais fornecidos pela PNS são para: (i) ponderar estimativas sobre características dos residentes na população; e (ii) ponderar características dos domicílios. O interesse deste trabalho focará no primeiro tipo de peso amostral.

O processo de cálculo dos pesos amostrais dos residentes na amostra da PNS 2013, descrito com maiores detalhes em Freitas (2014) e Souza-Júnior et al. (2015), é realizado em três etapas. Na primeira etapa, pesos amostrais base são definidos pelas recíprocas das probabilidades de inclusão na amostra das unidades em cada estágio do plano amostral. Na segunda etapa, os pesos bases passam por ajustes adicionais para compensar ausências de respostas. Na última etapa, são feitos ajustes para atender condições de calibração a valores populacionais conhecidos. Estas etapas podem ser descritas da seguinte forma:

1. A probabilidade de inclusão amostral da PNS do k -ésimo residente do j -ésimo domicílio do i -ésimo setor censitário no h -ésimo estrato da população pode ser expressa, de acordo com a estrutura do plano amostral descrita na Seção 2.1.1, por

$$\pi_{hijk} = \pi_{hi}\pi_{j|hi}\pi_{k|hij}, \quad (2.1)$$

em que:

- π_{hi} é a probabilidade de inclusão amostral do i -ésimo setor censitário do h -ésimo estrato;
- $\pi_{j|hi}$ é a probabilidade condicional de inclusão amostral do j -domicílio, dado que o i -ésimo setor censitário do h -ésimo estrato é selecionado para a amostra; e
- $\pi_{k|hij}$ é a probabilidade condicional de inclusão amostral do k -ésimo residente, dado que o j -domicílio do i -ésimo setor censitário do h -ésimo estrato é selecionado para a amostra.

Assim, pode-se expressar o peso amostral base do k -ésimo residente do j -ésimo domicílio i -ésimo setor censitário do h -ésimo estrato pela relação

$$d_{hijk} = \frac{1}{\pi_{hijk}} = \frac{1}{\pi_{hi}} \frac{1}{\pi_{j|hi}} \frac{1}{\pi_{k|hij}} \equiv d_{hi}d_{j|hi}d_{k|hij}, \quad (2.2)$$

em que $d_{hi} = 1/\pi_{hi}$, $d_{j|hi} = 1/\pi_{j|hi}$, $d_{k|hij} = 1/\pi_{k|hij}$ denotam pesos amostrais base para os três estágios de seleção da pesquisa. O cálculo destes pesos podem seguir o seguinte roteiro:

- a) pesos amostrais base do primeiro estágio (d_{hi}): suponha que m_h denota o número de setores censitários a serem selecionados do h -ésimo estrato no Cadastro Mestre para a Amostra Mestre. Similarmente, seja m_h^{PNS} denota o número de setores censitários selecionados da Amostra Mestre para a PNS. Dois aspectos do processo de seleção das unidades primárias no h -ésimo estrato são:
- amostragem dos m_h setores censitários do Cadastro Mestre para a Amostra Mestre foi feita por seleções independentes atribuindo-se probabilidade de seleção $p_{hi} = N_{hi}/N_h$ para o setor censitário (h, i) , em que N_{hi} denota o número de domicílios particulares ocupados (sem entrevistas conduzidas na época da seleção para a Amostra Mestre) do setor censitário (h, i) e N_h é o correspondente número de domicílios para o h -ésimo estrato.
 - amostragem aleatória simples dos m_h^{PNS} entre os m_h setores censitários da Amostra Mestre.

Assim, segue que a probabilidade de inclusão amostral do setor censitário (h, i) na PNS é a probabilidade deste setor ter sido selecionado do Cadastro Mestre para a Amostra Mestre ($m_h p_{hi}$) vezes a probabilidade dele ser selecionado também para a PNS dado que o setor foi selecionado para a Amostra Mestre (m_h^{PNS}/m_h), ou seja

$$\pi_{hi} = m_h p_{hi} \frac{m_h^{PNS}}{m_h} = p_{hi} m_h^{PNS}.$$

Dessa forma, os pesos amostrais base referentes ao primeiro estágio ficam dados por

$$d_{hi} = \frac{1}{\pi_{hi}} = \frac{1}{m_h} \frac{N_h}{N_{hi}} \frac{m_h}{m_h^{PNS}}, \quad (2.3)$$

para todo (h, i) na amostra da PNS.

- b) pesos amostrais do segundo estágio, condicionais as seleções do primeiro estágio ($w_{j|hi}$): sejam N_{hi}^* o número de domicílios ocupados e permanentes ocupados no setor censitário (h, i) e n_{hi} o tamanho da amostra de domicílios a ser selecionada no setor censitário (h, i) . Como as unidades do segundo estágio são, em cada estrato, selecionadas por amostragem aleatória simples, então

$$d_{j|hi} = \frac{1}{\pi_{j|hi}} = \frac{1}{n_{hi}/N_{hi}^*} = \frac{N_{hi}^*}{n_{hi}}, \quad (2.4)$$

para todo (h, i, j) .

- c) pesos amostrais do terceiro estágio, condicionais as seleções do primeiro e segundo estágios ($w_{k|hij}$): uma vez que o plano amostral do terceiro estágio é o de amostragem aleatória simples de um adulto entre os moradores elegíveis à pesquisa nos domicílios, segue então que

$$\pi_{k|hij} = \frac{1}{O_{hij}},$$

em que O_{hij} representa o número de moradores com 18 ou mais no domicílio (h, i, j) da população alvo. Neste caso, os pesos amostrais do terceiro estágio, condicionais as seleções dos estágios anteriores são

$$d_{k|hij} = \frac{1}{\pi_{k|hij}} = O_{hij}, \quad (2.5)$$

para todo (h, i, j, k) .

Portanto, os pesos amostrais base para os residentes selecionados pela PNS são por (2.2) dados pelo produto de (2.3)–(2.5), ou seja

$$d_{hijk} = \frac{1}{m_h} \frac{N_h}{N_{hi}} \frac{m_h}{m_h^{PNS}} \frac{N_{hi}^*}{n_{hi}} O_{hij}, \quad (2.6)$$

para todo (h, i, j, k) .

2. Ajustes para não-respostas na PNS precisam compensar perdas de entrevistas sobre características dos domicílios, assim como perdas nas entrevistas individuais sobre os residentes selecionados no terceiro estágio da pesquisa. Uma descrição para estes ajustes são:

- a) correção para a perda de entrevistas sobre informações domiciliares: utiliza-se o método da ponderação dentro de classes (LOHR, 2010, p. 340–342), em que as classes são dadas por cada setor censitário de cada estrato e o fator de ponderação da classe corresponde ao inverso da taxa de resposta (entrevista completa) entre todos os domicílios da classe. Os correspondentes pesos amostrais do segundo estágio são, após o ajuste, dados por

$$d_{j|hi}^* = \frac{1}{\pi_{j|hi} \widehat{\phi}_{hij}} = \frac{N_{hi}^* n_{hi}^*}{n_{hi} n_{hi}^{**}}, \quad (2.7)$$

para todo (h, i, j) , em que n_{hi}^* denota o número de entrevistas domiciliares conduzidas dentre os n_{hi}^{**} domicílios selecionados (ocupados e com residentes) do setor censitário (h, i) e $\widehat{\phi}_{hij} = n_{hi}^{**}/n_{hi}^*$ é a probabilidade estimada de entrevista no domicílio (h, i, j) ser completa.

- b) correção para perda das entrevistas sobre os indivíduos selecionados: também é feita aplicando-se o método da ponderação dentro de classes. Para a formação das classes, observou-se que as perdas nas entrevistas ocorreram mais frequentemente entre moradores do sexo masculino do que nos moradores do sexo feminino. Portanto, as classes foram dadas pelos dois grupos de sexo e taxas de entrevistas completas dentro destes dois grupos são usadas para obter as probabilidades estimadas de entrevista do indivíduo selecionado ser completa, denotada por $\widehat{\delta}_{hijk}^M$ se o residente (h, i, j, k) é do sexo masculino e $\widehat{\delta}_{hijk}^F$ se o residente (h, i, j, k) é do sexo feminino, para todo (h, i, j, k) . Os pesos amostrais ajustados para o terceiro estágio são dados por

$$d_{k|hij}^* = \frac{1}{\pi_{k|hij} \widehat{\delta}_{hijk}} = \frac{O_{hij}}{\widehat{\delta}_{hijk}}, \quad (2.8)$$

para todo (h, i, j, k) , em que

$$\widehat{\delta}_{hijk} = \widehat{\delta}_{hijk}^M \alpha_{hijk}^M + \widehat{\delta}_{hijk}^F \alpha_{hijk}^F,$$

$\alpha_{hijk}^M = 1$ se o residente (h, i, j, k) é do sexo masculino e $\alpha_{hijk}^M = 0$ caso contrário, $\alpha_{hijk}^F = 1$ se o residente (h, i, j, k) é do sexo feminino e $\alpha_{hijk}^F = 0$ caso contrário. Assim, usando (2.3), (2.7) e (2.8), os pesos amostrais após os ajustes para não respostas dos residentes do sexo masculino ou feminino ficam dados respectivamente por

$$d_{hijk}^* = d_{hi} d_{j|hi}^* d_{k|hij}^* = \frac{1}{m_h} \frac{N_h}{N_{hi}} \frac{m_h}{m_h^{PNS}} \frac{N_{hi}^* n_{hi}^*}{n_{hi} n_{hi}^{**}} \frac{O_{hij}}{\widehat{\delta}_{hijk}}, \quad (2.9)$$

para todo (h, i, j, k) .

3. Ajustes para calibração na PNS foram planejados para igualar certas estimativas da amostra dos residentes (amostra menor), em que apenas um desses residentes foi selecionado aleatoriamente em cada domicílio da amostra, com as estimativas dos dados coletados na amostra de domicílios (amostra maior), que obteve informações sobre todos os residentes de cada domicílio. Os pesos amostrais ajustados dos residentes $\{d_{hijk}^*\}$, em (2.9), foram calibrados para que as estimativas de totais populacionais por sexo e pelas faixas etárias: 18 a 24, 25 a 39, 40 a 59 e 60 anos ou mais correspondesse às estimativas obtidas nestes grupos com a amostra maior. Os pesos calibrados para os residentes da amostra são

$$w_{hijk}^* = d_{hijk}^* \left(\frac{\widehat{N}_{a,m,c}^D}{\widehat{N}_{a,m,c}^R} \alpha_{hijk}^M + \frac{\widehat{N}_{a,f,c}^D}{\widehat{N}_{a,f,c}^R} \alpha_{hijk}^F \right), \quad (2.10)$$

em que

- $\widehat{N}_{a,m,c}^D$ é a estimativa do número populacional de residentes do sexo masculino na c -ésima faixa etária e nível geográfico a , obtida usando a amostra de domicílios ($c = 1, 2, 3, 4$);
- $\widehat{N}_{a,m,c}^R$ é a estimativa do número populacional de residentes do sexo masculino na c -ésima faixa etária obtida usando a amostra de respondentes ($c = 1, 2, 3, 4$);
- $\widehat{N}_{a,f,c}^D$ é a estimativa do número populacional de residentes do sexo feminino na c -ésima faixa etária obtida usando a amostra de domicílios ($c = 1, 2, 3, 4$);
- $\widehat{N}_{a,f,c}^R$ é a estimativa do número populacional de residentes do sexo feminino na c -ésima faixa etária obtida usando a amostra de respondentes ($c = 1, 2, 3, 4$).

Os pesos amostrais em (2.10) são, portanto, os pesos amostrais finais da PNS.

2.2 Métodos de estimação

2.2.1 Estimação de totais e médias

Os pesos amostrais finais

$$\{w_{hijk}^* : (h, i, j, k) \text{ é unidade respondente na PNS}\}, \quad (2.11)$$

dados em (2.10), são construídos de forma que uma estimativa válida do total populacional de uma dada variável y possa ser calculada usando o estimador ponderado

$$\widehat{t}_y = \sum_{h=1}^H \widehat{t}_{yh}, \quad (2.12)$$

em que

$$\hat{t}_{yh} = \sum_i \sum_j \sum_k w_{hijk}^* y_{hijk}$$

é o estimador do total da variável y no h -ésimo estrato, y_{hijk} denota o valor de y para o residente (h, i, j, k) e os somatórios nos índices i, j e k denotam somas sobre os valores possíveis destes índices na amostra.

Um método alternativo de estimação que facilita a análise de dados amostrais complexos é o do conglomerado primário (HANSEN; HURWITZ; MADOW, 1953a; HANSEN; HURWITZ; MADOW, 1953b; DURBIN, 1953). Neste método, os residentes de todos os domicílios de um mesmo setor censitário são aglomerados para formar um único conglomerado. Dessa forma, pode-se reescrever os pesos amostrais (2.11) da seguinte forma:

$$\{w_{hik}^* : h = 1, \dots, H; i = 1, \dots, m_h; k = 1, \dots, m_{hi}\}, \quad (2.13)$$

em que w_{hik}^* denota o peso amostral do k -ésimo residente do i -ésimo conglomerado primário do h -ésimo estrato, m_h é o número de conglomerados primários no estrato h e m_{hi} denota, de agora em diante, o número de residentes no i -ésimo conglomerado primário do h -ésimo estrato. Assim, pode-se reescrever o estimador do total populacional de uma dada variável y em (2.12) por

$$\hat{t}_y = \sum_{h=1}^H \hat{t}_{yh}, \quad (2.14)$$

em que

$$\hat{t}_{yh} = \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^* y_{hik}$$

e y_{hik} é o valor de y para o k -ésimo residente do i -ésimo conglomerado primário do h -ésimo estrato.

O estimador ponderado para a média populacional de y fica dado por

$$\bar{y}_w = \frac{\hat{t}_y}{\hat{N}}, \quad (2.15)$$

em que

$$\hat{N} = \sum_{h=1}^H \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^*$$

é o estimador ponderado do número de residentes da população alvo.

2.2.2 Estimação em domínios

Comumente, deseja-se produzir estimativas na pesquisa de totais ou médias populacionais de y não somente para toda a população alvo de residentes, mas também estimativas destes parâmetros em diferentes subpopulações de residentes, denominadas de domínios. Este

tipo de estimação é bastante facilitada com o uso de variáveis indicadoras. Sejam S um determinado domínio de interesse e

$$I_{S,hik} = \begin{cases} 1, & \text{se o residente } (h, i, k) \text{ pertence ao domínio } S, \\ 0, & \text{caso contrário.} \end{cases}$$

Seguindo Heeringa, West e Berglund (2010, p. 112), os correspondentes estimadores do total e da média populacional de y no domínio S ficam dados por

$$\hat{t}_{y,S} = \sum_{h=1}^H \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^* I_{S,hik} y_{hik} \quad (2.16)$$

e

$$\bar{y}_{w,S} = \frac{\hat{t}_{y,S}}{\hat{N}_S}, \quad (2.17)$$

em que

$$\hat{N}_S = \sum_{h=1}^H \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^* I_{S,hik}.$$

2.2.3 Estimação de frequências e proporções

Os estimadores (2.14) e (2.15) são definidos para variáveis y gerais, pondendo elas serem do tipo contínua, discreta ou categorizada. No caso de y ser uma variável discreta ou categorizada, suponha sem perda de generalidade que ela assuma os valores $1, \dots, C$. Um parâmetro que é frequentemente de interesse neste caso é a distribuição de frequências dos valores de y entre os residentes da população, denotada aqui pelo vetor $(N_1, \dots, N_C)^\top$. Outro parâmetro relacionado é a distribuição das proporções dos diferentes valores de y na população, dada pelo vetor $(p_1, p_2, \dots, p_C)^\top$.

O estimador do vetor de frequências populacionais pode ser obtido por (2.12) através do vetor de frequências estimadas

$$(\hat{N}_1, \dots, \hat{N}_C)^\top, \quad (2.18)$$

em que

$$\hat{N}_c = \sum_{h=1}^H \sum_i \sum_j \sum_k w_{hijk}^* I(y_{hijk} = c)$$

para todo $c = 1, \dots, C$ e $I(A) = 1$ se o evento A ocorre e $I(A) = 0$ caso contrário. Similarmente, usando (2.15), o estimador do vetor de proporções dos valores de y na população é

$$(\hat{p}_1, \hat{p}_2, \dots, \hat{p}_C)^\top, \quad (2.19)$$

em que

$$\hat{p}_c = \frac{\hat{N}_c}{\hat{N}} = \frac{1}{\hat{N}} \sum_{h=1}^H \sum_i \sum_j \sum_k w_{hijk}^* I(y_{hijk} = c),$$

para todo $c = 1, \dots, C$.

Com a inclusão dos indicadores de presença em domínios, conforme (2.16) e (2.17), pode-se estender as definições dos vetores estimadores (2.18) e (2.19) para produzir estimativas em domínios.

2.2.4 Estimação de variâncias

A estimação das variâncias e covariâncias dos estimadores nas Seções 2.2.1–2.2.3 apresenta algumas dificuldades em razão da complexidade do plano amostral da pesquisa e dos estimadores propostos serem funções não lineares. As dificuldades em relação aos aspectos do plano amostral podem ser contornadas adotando-se também o procedimento simplificado para estimar as variâncias que se baseia na aproximação pelo conglomerado primário, descrito na Seção 2.2.1.

O problema da não linearidade dos estimadores pode ser resolvido pela técnica da Linearização de Taylor (WOLTER, 2007, Cap. 6). Nesta técnica, os estimadores são aproximados por funções lineares e a variância destas aproximações é usada para estimar a variância do estimador. Por exemplo, para estimar a variância do estimador \bar{y}_w em (2.15) usando a linearização de Taylor considera-se, primeiro, a aproximação

$$\begin{aligned} \text{var}(\bar{y}_w) &\approx \text{var} \left\{ \bar{y}_U + \frac{1}{N} (\hat{t}_y - \bar{y}_U \hat{N}) \right\} \\ &= \frac{1}{N^2} \left\{ \text{var}(\hat{t}_y) + \bar{y}_U^2 \text{var}(\hat{N}) - 2\bar{y}_U \text{cov}(\hat{t}_y, \hat{N}) \right\}, \end{aligned}$$

em que $\bar{y}_U = t_y/N$ e N denotam a média de y e o número de elementos na população. Em seguida, de posse das estimativas das variâncias $\hat{v}(\hat{t}_y)$ e $\hat{v}(\hat{N})$ e da estimativa da covariância dos totais $\hat{cov}(\hat{t}_y, \hat{N})$ dos estimadores \hat{t}_y e \hat{N} , segue que a estimativa pela linearização de Taylor da variância de \bar{y}_w é

$$\hat{v}(\bar{y}_w) = \frac{1}{\hat{N}^2} \left\{ \hat{v}(\hat{t}_y) + \bar{y}_w^2 \hat{v}(\hat{N}) - 2\bar{y}_w \hat{cov}(\hat{t}_y, \hat{N}) \right\}.$$

O cálculo de variâncias estimadas usando esta abordagem será feito aqui neste trabalho com o uso do pacote `survey` (LUMLEY, 2011) do software R.

Com base na estimativa $\hat{v}(\hat{\theta})$ da variância de um determinado estimador $\hat{\theta}$ sob o plano amostral adotado – seja este estimador um total, média ou proporção – três medidas estimadas frequentes usadas para avaliar a precisão das estimativas são:

- erro padrão de $\hat{\theta}$:

$$EP(\hat{\theta}) = \sqrt{\hat{v}(\hat{\theta})};$$

- coeficiente de variação de $\hat{\theta}$:

$$CV(\hat{\theta}) = \frac{EP(\hat{\theta})}{\theta};$$

- efeito do plano amostral (EPA) na estimação de $\hat{\theta}$:

$$EPA(\hat{\theta}) = \frac{\hat{v}(\hat{\theta})}{\hat{v}_{AAS}(\hat{\theta})},$$

em que $\hat{v}_{AAS}(\hat{\theta})$ denota a variância de $\hat{\theta}$ sob um plano de amostragem aleatória simples com o mesmo número de elementos que o plano adotado.

2.3 Teste de independência

Sejam x e y duas variáveis categorizadas de interesse, assumindo os valores $1, \dots, R$ e $1, \dots, C$, respectivamente. Suponha que as respostas x e y de cada residente da população são classificadas simultaneamente. A distribuição de frequências dos pares de valores possíveis, $\{N_{rc} : r = 1 \dots, R; c = 1, \dots, C\}$, é dada na Tabela 1. Também, nesta tabela, são apresentadas as distribuições de frequências marginais populacionais de x e y , denotadas respectivamente por $\{N_{r+} = \sum_{c=1}^C N_{rc} : r = 1 \dots, R\}$ e $\{N_{+c} = \sum_{r=1}^R N_{rc} : c = 1, \dots, C\}$. Portanto, por definição, as frequências conjuntas e marginais satisfazem as relações

$$\sum_{r=1}^R \sum_{c=1}^C N_{rc} = \sum_{r=1}^R N_{r+} = \sum_{c=1}^C N_{+c} = N.$$

Tabela 1 – Tabela de contingência populacional das variáveis x e y

x	y				Total
	1	2	\dots	C	
1	N_{11}	N_{12}	\dots	N_{1C}	N_{1+}
2	N_{21}	N_{22}	\dots	N_{2C}	N_{2+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
R	N_{R1}	N_{R2}	\dots	N_{RC}	N_{R+}
Total	N_{+1}	N_{+2}	\dots	N_{+C}	N

A hipótese nula estatística de independência entre x e y na população pode ser expressa por

$$H_0 : p_{rc} = p_{r+}p_{+c}, \quad \text{para todas as células } (r, c) \text{ da Tabela 1,} \quad (2.20)$$

em que $p_{rc} = N_{rc}/N$ denota a proporção populacional da célula (r, c) , $p_{r+} = N_{r+}/N$ é a proporção marginal da linha r e $p_{+c} = N_{+c}/N$ é a proporção marginal da coluna c . Uma

forma equivalente para a hipótese (2.20) é

$$H_0 : \mathbf{F} = \mathbf{0}, \quad (2.21)$$

em que

$$\mathbf{F} = (F_{11}, \dots, F_{1,C-1}, \dots, F_{R-1,1}, \dots, F_{R-1,C-1})^\top,$$

$F_{rc} = p_{rc} - p_{r+}p_{+c}$ ($r = 1, \dots, R-1$ e $c = 1, \dots, C-1$) e $\mathbf{0}$ é vetor $(R-1)(C-1) \times 1$ formado por zeros.

Para formular um teste para a hipótese nula de independência (2.21), incorporando-se as características do plano amostral, pode-se utilizar as seguintes etapas (LEHTONEN; PAHKINEN, 2004):

1. Estimação das frequências populacionais: note que as componentes do vetor \mathbf{F} são funções das proporções populacionais $\{p_{rc} : r = 1, \dots, R-1; c = 1, \dots, C-1\}$. Para estimar estas proporções, considere primeiro a estimacão das frequências populacionais das células através do estimador

$$\hat{n}_{rc} = \sum_{h=1}^H \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^* \left(\frac{n}{\hat{N}} \right) y_{hik,rc}, \quad (2.22)$$

em que

$$y_{hik,rc} = \begin{cases} 1, & \text{se o residente } (h, i, k) \text{ está na célula } (r, c), \\ 0, & \text{caso contrário,} \end{cases}$$

$n = \sum_{h=1}^H \sum_{i=1}^{m_h} m_{hi}$ é o tamanho da amostra e \hat{N} é dado em (2.15). A constante n/\hat{N} em (2.22) é um fator de normalização para garantir que

$$\sum_{r=1}^R \sum_{c=1}^C \hat{n}_{rc} = n.$$

As frequências ponderadas estimadas \hat{n}_{rc} podem ser organizadas como na Tabela 2, juntamente com as correspondentes frequências marginais estimadas. Assim, a

Tabela 2 – Tabela de contingência estimada ponderada das variáveis x e y

x	y				Total
	1	2	...	C	
1	\hat{n}_{11}	\hat{n}_{12}	...	\hat{n}_{1C}	\hat{n}_{1+}
2	\hat{n}_{21}	\hat{n}_{22}	...	\hat{n}_{2C}	\hat{n}_{2+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
R	\hat{n}_{R1}	\hat{n}_{R2}	...	\hat{n}_{RC}	\hat{n}_{R+}
Total	\hat{n}_{+1}	\hat{n}_{+2}	...	\hat{n}_{+C}	$\hat{n}_{++} = n$

proporção estimada de residentes na célula (r, c) fica dada por

$$\hat{p}_{rc} = \frac{\hat{n}_{rc}}{n} = \frac{1}{n} \sum_{h=1}^H \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^* \left(\frac{n}{\widehat{N}} \right) y_{hik,rc} = \frac{1}{\widehat{N}} \sum_{h=1}^H \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^* y_{hik,rc}, \quad (2.23)$$

$(r = 1, \dots, R; c = 1, \dots, C)$ e as proporções marginais estimadas são

$$\hat{p}_{r+} = \sum_{c=1}^C \hat{p}_{rc} = \frac{\hat{n}_{r+}}{n} = \frac{1}{\widehat{N}} \sum_{h=1}^H \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^* y_{hik,r}. \quad (r = 1, \dots, R) \quad (2.24)$$

e

$$\hat{p}_{+c} = \sum_{r=1}^R \hat{p}_{rc} = \frac{\hat{n}_{+c}}{n} = \frac{1}{\widehat{N}} \sum_{h=1}^H \sum_{i=1}^{m_h} \sum_{k=1}^{m_{hi}} w_{hik}^* y_{hik,c} \quad (c = 1, \dots, C), \quad (2.25)$$

em que

$$y_{hik,r} = \begin{cases} 1, & \text{se o residente } (h, i, k) \text{ está na } r\text{-ésima linha da tabela,} \\ 0, & \text{caso contrário,} \end{cases}$$

e

$$y_{hik,c} = \begin{cases} 1, & \text{se o residente } (h, i, k) \text{ está na } c\text{-ésima coluna da tabela,} \\ 0, & \text{caso contrário.} \end{cases}$$

Os estimadores (2.23), (2.24) e (2.25) correspondem a estimadores ponderados de p_{rc} , p_{r+} e p_{+c} usando os pesos amostrais da pesquisa. Segue, portanto, que o estimador do vetor \mathbf{F} em (2.21) é dado por

$$\widehat{\mathbf{F}} = (\widehat{F}_{11}, \dots, \widehat{F}_{1,C-1}, \dots, \widehat{F}_{R-1,1}, \dots, \widehat{F}_{R-1,C-1})^\top,$$

em que $\widehat{F}_{rc} = \hat{p}_{rc} - \hat{p}_{r+}\hat{p}_{+c}$ ($r = 1, \dots, R-1$ e $c = 1, \dots, C-1$).

2. Estatística de Pearson (não corrigida): usando as frequências da Tabela 2, obtém-se a estatística de Pearson para testar H_0 em (2.21) pela expressão

$$X_P^2 = \sum_{r=1}^R \sum_{c=1}^C \frac{(\hat{n}_{rc} - \hat{n}_{r+}\hat{n}_{+c})^2}{\hat{n}_{r+}\hat{n}_{+c}} = n \sum_{r=1}^R \sum_{c=1}^C \frac{(\hat{p}_{rc} - \hat{p}_{r+}\hat{p}_{+c})^2}{\hat{p}_{r+}\hat{p}_{+c}} = n \widehat{\mathbf{F}}^\top \widehat{\mathbf{P}}_{0F}^{-1} \widehat{\mathbf{F}}, \quad (2.26)$$

em que

$$\widehat{\mathbf{P}}_{0F} = \widehat{\mathbf{H}}^\top \widehat{\mathbf{P}}_0 \widehat{\mathbf{H}},$$

$\widehat{\mathbf{H}}$ é matriz de dimensão $(R-1)(C-1) \times (R-1)(C-1)$ de derivadas parciais de \mathbf{F} com respeito a p_{rc} avaliada em $p_{rc} = \hat{p}_{rc}$,

$$\widehat{\mathbf{P}}_0 = \text{diag}(\widehat{\mathbf{p}}_0) - \widehat{\mathbf{p}}_0 \widehat{\mathbf{p}}_0^\top,$$

$$\widehat{\mathbf{p}}_0 = (\widehat{p}_{0,11}, \dots, \widehat{p}_{0,1,C-1}, \dots, \widehat{p}_{0,R-1,1}, \dots, \widehat{p}_{0,R-1,C-1})^\top,$$

$\widehat{p}_{0,rc} = \hat{p}_{r+}\hat{p}_{+c}$ ($r = 1, \dots, R-1$ e $c = 1, \dots, C-1$) e $\text{diag}(\widehat{\mathbf{p}}_0)$ é a matriz diagonal com elementos diagonais dados pelas componentes de $\widehat{\mathbf{p}}_0$.

3. Correção de Rao-Scott de primeira ordem: sejam $\{\widehat{d}_{rc}\}$ os efeitos do plano amostral das proporções estimadas $\{\widehat{p}_{rc}\}$ e, similarmente, $\{\widehat{d}_{r+}\}$ e $\{\widehat{d}_{+c}\}$ os correspondentes efeitos marginais. A correção de Rao-Scott de primeira ordem para X_P^2 usa a estatística

$$X_{RS,1}^2 = \frac{X_P^2}{\widehat{\delta}}, \quad (2.27)$$

em que

$$\widehat{\delta} = \frac{1}{(R-1)(C-1)} \left\{ \sum_{r=1}^R \sum_{c=1}^C (1 - \widehat{p}_{rc}) \widehat{d}_{rc} - \sum_{r=1}^R (1 - \widehat{p}_{r+}) \widehat{d}_{r+} - \sum_{c=1}^C (1 - \widehat{p}_{+c}) \widehat{d}_{+c} \right\}.$$

O teste da hipótese nula é feito referindo-se o valor da estatística $X_{RS,1}^2$ com a distribuição qui-quadrado com $(R-1)(C-1)$ graus de liberdade. Um teste alternativo com correção de primeira ordem compara o valor da estatística F corrigida

$$F_{RS,1} = \frac{X_{RS,1}^2}{(R-1)(C-1)} \quad (2.28)$$

com os quantis da distribuição F com graus de liberdade $(R-1)(C-1)$ e $(R-1)(C-1)\kappa$, em que κ corresponde aos graus de liberdade que o plano amostral atribui para estimação de variâncias.

4. Correção de Rao-Scott de segunda ordem: seja a matriz de efeitos do planos generalizados dada por

$$\widehat{\mathbf{D}} = n \widehat{\mathbf{P}}_{0F}^{-1} \widehat{\mathbf{V}}_F,$$

em que $\widehat{\mathbf{P}}_{0F}$ é definido em (2.26),

$$\widehat{\mathbf{V}}_F = \widehat{\mathbf{H}}^\top \widehat{\mathbf{V}}_{des} \widehat{\mathbf{H}}$$

e $\widehat{\mathbf{V}}_{des}$ é um estimador consistente da matriz de variância-covariância do vetor de proporções estimadas

$$\widehat{\mathbf{p}} = (\widehat{p}_{11}, \dots, \widehat{p}_{1,C-1}, \dots, \widehat{p}_{R-1,1}, \dots, \widehat{p}_{R-1,C-1})^\top.$$

A correção de segunda ordem para a estatística X_P^2 é dada pela estatística corrigida

$$X_{RS,2}^2 \equiv X_P^2(\widehat{\delta}, \widehat{a}^2) = \frac{X_P^2}{\widehat{\delta}(1 + \widehat{a}^2)}, \quad (2.29)$$

em que

$$\widehat{\delta} = (\widehat{\delta}_1, \dots, \widehat{\delta}_{(R-1)(C-1)})^\top = \frac{\text{tr}(\widehat{\mathbf{D}})}{(R-1)(C-1)}$$

e

$$\widehat{a}^2 = \sum_{\ell=1}^{(R-1)(C-1)} \frac{\widehat{\delta}_\ell^2}{(R-1)(C-1)\widehat{\delta}^2} - 1.$$

Sob H_0 , a distribuição de $X_P^2(\hat{\delta}, \hat{a}^2)$ é assintoticamente qui-quadrado com graus de liberdade (ajustados pelo método de Satterthwaite)

$$\nu = \frac{(R-1)(C-1)}{1 + \hat{a}^2}.$$

Uma melhor aproximação pode ser obtida pela correção de Rao-Scott de segunda ordem para a estatística F que é dada por

$$F_{RS,2} = \frac{X_{RS,2}^2}{(R-1)(C-1)}, \quad (2.30)$$

em que $X_{RS,2}^2$ é a estatística de Pearson com correção de segunda ordem dada em (2.29), cuja distribuição nula aproximada é a distribuição F com graus de liberdade ν e $\nu\kappa$.

3 Análise dos resultados

Neste capítulo, serão apresentadas as análises dos dados do trabalho. Todos os resultados foram obtidos com a ajuda do software R (R Core Team, 2018), versão 3.5.1. O processo de estimação de parâmetros e testes de hipótese de interesse foram realizados utilizando-se o pacote `survey` (LUMLEY, 2011) do R. Entre as análises realizadas, examina-se em primeiro lugar características das distribuições univariadas das variáveis de interesse. Em seguida, discute-se o estudo das associações entre a autoavaliação da saúde com cada uma das variáveis socio-demográficas, isto é sexo, faixa etária, raça, escolaridade e situação habitacional. Os testes de associação empregam as correções de Rao-Scott de segunda ordem para as estatísticas de Pearson e F , dadas em (2.29) e (2.30), respectivamente.

3.1 Análise das distribuições univariadas

De acordo com a Tabela 3, os pesos amostrais dos moradores da PNS 2013, referente à região Nordeste, apresentam considerável variação entre eles. A correspondente distribuição é assimétrica à direita, com uma variação de 27,57 a 76.207,38 e mediana igual a 1.276,30. A soma destes pesos é 146.308.458,00, que representa uma estimativa, no momento da pesquisa, do número de moradores com 18 anos ou mais na população alvo do país.

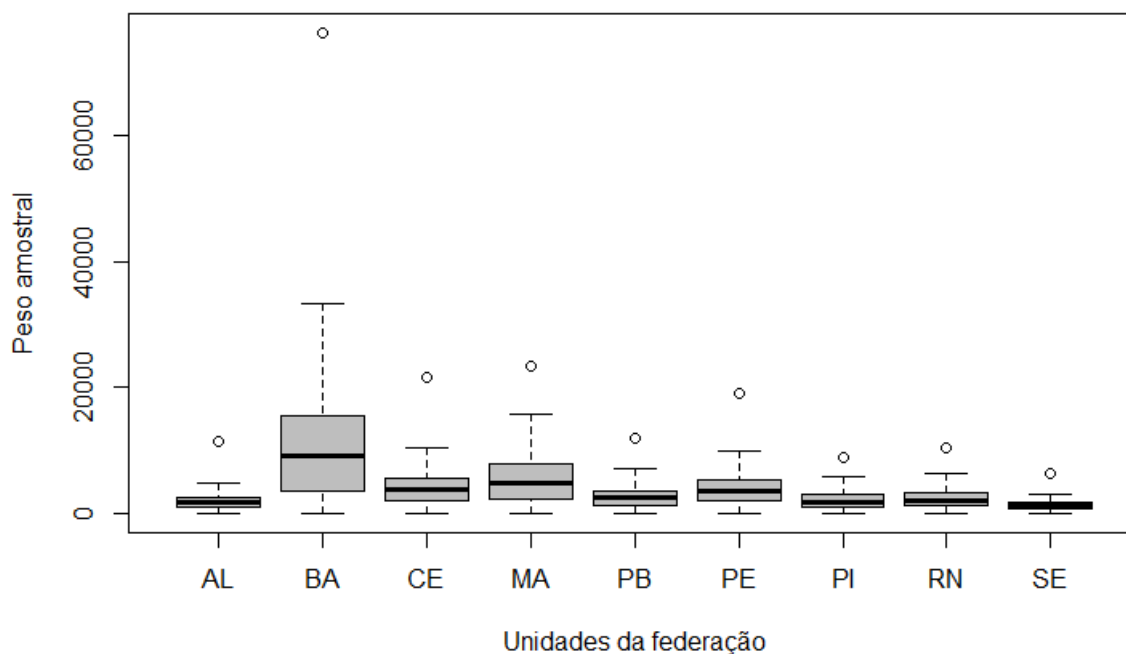
Tabela 3 – Estatísticas descritivas para os pesos amostrais dos moradores com correções para não resposta e calibração, Região Nordeste, 2013.

Estatística	Valor
Mínimo	27,57
Média	2.127,70
Máximo	76.207,38
Soma	146.308.458,00
Percentis	
1%	203,62
5%	328,56
25%	704,18
50%	1.276,30
75%	2.445,48
95%	6.506,99
99%	13.894,79

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Os diagramas de caixa destes pesos amostrais exibido na Figura 1 indicam distribuições aproximadamente homogêneas entre os estados nordestinos, exceto no estado da Bahia, que possui maior variabilidade interquartílica e ocorrência de um peso amostral com valor superior a 60.000.

Figura 1 – Diagramas de caixa dos pesos amostrais dos moradores com correções para não respostas e calibração por Unidade Federativa da Região Nordeste, 2013.



Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Tamanhos amostrais e estimativas para as proporções de indivíduos na população, sem correção (proporção amostral) e com correção pelo plano amostral (proporção ponderada pelos pesos amostrais), para as categorias das variáveis autoavaliação da saúde, sexo, faixa etária, raça, escolaridade e situação habitacional são dadas na Tabela 4. Esta tabela apresenta também estimativas para os coeficientes de variação e os efeitos do plano amostral para cada proporção estimada. O tamanho total da amostra na base de dados é 18.305 respondentes.

Com base nas proporções ponderadas pelos pesos amostrais para as categorias da variável autoavaliação da saúde, dadas na Tabela 4, estima-se que cerca de 56,7% dos moradores adultos da população alvo possuem avaliação da saúde boa ou muito boa. Estima-se, também, que 35,0% destes moradores tem autoavaliação regular e 8,3% autoavaliação ruim ou muito ruim. Estas estimativas ponderadas não apresentam grandes diferenças das correspondentes estimativas baseadas nas proporções amostrais, que não incorporam as informações do plano amostral, ou seja 58,3%; 33,3% e 8,4%, respectivamente.

Tabela 4 – Número de entrevistados (n), proporção amostral, proporção ponderada, coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias das variáveis da análise, Região Nordeste, 2013.

Variável	Categoria	n	Proporção Amostral (%)	Proporção ponderada (%)	CV (%)	EPA
Autoavaliação da saúde	Boa	10.664	58,3	56,7	1,0	2,5
	Regular	6.102	33,3	35,0	1,6	2,6
	Ruim	1.539	8,4	8,3	3,8	2,4
Sexo	Feminino	10.545	57,6	53,1	0,0	0,0
	Masculino	7.760	42,4	46,9	0,0	0,0
Faixa etária	18-29	4.457	24,3	28,2	1,0	0,7
	30-59	10.454	57,1	54,7	0,5	0,6
	60-64	1.035	5,7	5,4	4,0	1,6
	65-74	1.463	8,0	7,3	3,2	1,5
	75+	896	4,9	4,5	4,5	1,8
Raça	Amarela ou Indígena	269	1,5	1,5	9,5	2,4
	Branca	5.090	27,8	26,9	2,4	3,9
	Parda	10.830	59,2	59,6	1,3	4,4
	Preta	2.116	11,6	12,0	4,0	3,9
Escolaridade	Sem instrução ou fundamental incompleto	8.532	46,6	48,1	1,4	3,2
	Fundamental completo ou médio incompleto	2.621	14,3	14,4	3,1	2,9
	Médio completo ou superior incompleto	5.408	29,5	29,3	2,0	3,0
	Superior completo	1.744	9,5	8,3	4,3	3,0
	Situação habitacional	Rural	4.228	23,1	23,8	2,2
	Urbano	14.077	76,9	76,2	0,7	2,6

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Com respeito às proporções de indivíduos da população alvo referentes às características sócio-demográficas, estima-se que dos moradores da população adulta do Nordeste em 2013, aproximadamente:

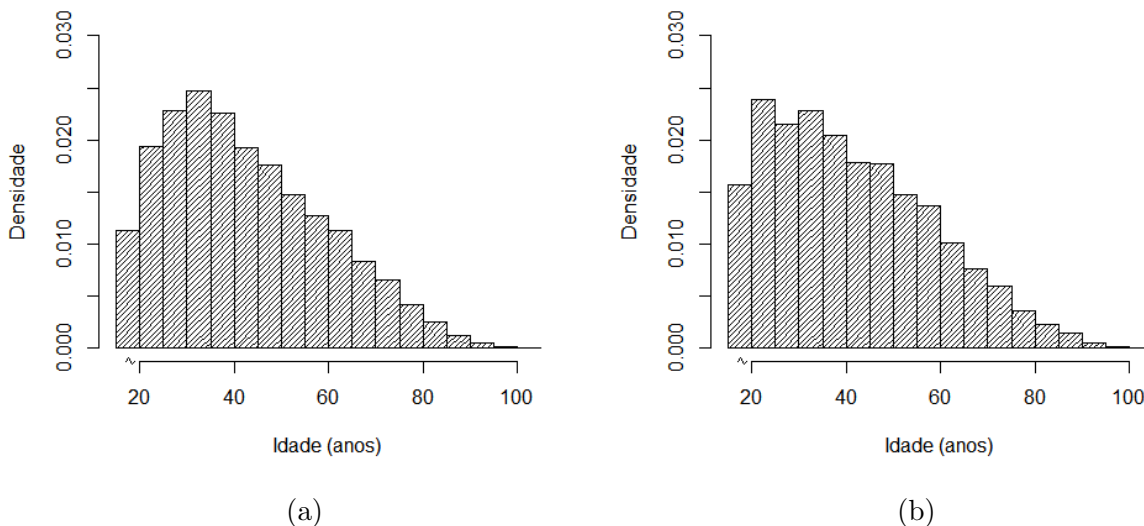
- 53,1% são do sexo feminino;
- 82,9% dos moradores têm entre 18 a 59 anos de idade;
- 59,6% referem a cor parda, 26,9% a branca, 12,0% preta e 1,5% amarela ou indígena;
- 62,5% possuem escolaridade inferior ao ensino médio, 29,3% possuem ensino médio completo sem ensino superior completo e 8,3% possuem ensino superior completo;
- 76,2% dos domicílios de residência dos moradores eram em locais urbanos.

As estimativas ponderadas para estas características sócio-demográficas apresentam, em alguns casos, diferenças maiores para as respectivas proporções não ponderadas do que é observado na comparação destas estimativas para as categorias da variável autoavaliação da saúde. Por exemplo, sem levar em consideração o plano amostral da pesquisa, a

porcentagem de moradores do sexo feminino é estimada em 57,6%, que apresenta 4,5% pontos percentuais a mais que a correspondente proporção ponderada. Outro caso é na categoria de moradores com 18 a 29 anos, em que a porcentagem de moradores sem ponderação é 3,9% menor que a porcentagem ponderada estimada da mesma categoria.

Em virtude da variável idade apresentar informação quantitativa na base de dados, obteve-se o histograma mostrado na Figura 2 para visualizar melhor a distribuição desta variável na população. Neste gráfico, as densidades de cada classe são estimadas com as proporções amostrais da classe e com as proporções ponderadas pelos pesos amostrais. Nota-se, em ambos os casos, que as distribuições estimadas das idades dos moradores na população são assimétricas à direita, com maior concentração de moradores até os 50 anos de idade. Contudo, a distribuição com correção usa estimativas não-tendenciosas sendo, portanto, preferível neste caso. Claramente, observa-se uma subestimação das estimativas não corrigidas para idades até 25 anos e uma superestimação, por exemplo, entre os moradores com 30 e 35 anos.

Figura 2 – Histogramas das idades dos moradores da região Nordeste - PNS 2013. Em (a), as densidades são estimadas usando as proporções amostrais de indivíduos (sem correção) em cada classe etária e, em (b), as densidades são estimadas com as estimativas das proporções ponderadas pelos pesos amostrais, Região Norteste, 2013.



Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Os efeitos do plano amostral (EPA) dados na Tabela 4 indicam diferentes situações para comparar a eficiência de estimação sob o plano amostral adotado pela PNS em relação a uma amostra aleatória simples com o mesmo número de moradores. Para as proporções estimadas da avaliação de saúde pelo plano amostral da PNS 2013, os efeitos indicam que as variâncias dessas proporções são, em média, 2,5 vezes as respectivas variâncias que seriam obtidas se o plano amostral da pesquisa correspondesse a uma amostragem aleatória simples com a mesma quantidade de moradores respondentes. Para as proporções associadas às características sócio-demográficas, os efeitos variam entre 0 e 4,4. Nas categorias da variável sexo e faixas etárias 19 a 29 e 30 a 59 anos, os efeitos do plano amostral são inferiores a 1, sugerindo uma maior eficiência do plano amostral da PNS. Nas demais categorias, os efeitos são superiores a 1, indicando perda de eficiência do plano amostral da PNS. O maior efeito ocorre na estimação da proporção de moradores referindo-se à cor parda.

3.2 Estudo de associações

3.2.1 Associação entre autoavaliação de saúde e sexo

A distribuição conjunta das porcentagens estimadas com a ponderação pelos pesos amostrais, segundo o sexo e a autoavaliação da saúde dos moradores adultos, é dada na Tabela 5, enquanto que as distribuições das porcentagens estimadas de moradores adultos da autoavaliação da saúde condicional ao sexo são dadas na Tabela 6.

De acordo com a distribuição conjunta da Tabela 5, nota-se que as porcentagens mais elevadas ocorrem nas categorias: masculino e boa/muito boa (28,9%) e feminino e boa/muito boa (27,8%); as maiores porcentagens seguintes referem-se às categorias de autoavaliação regular para os grupos de sexo feminino e masculino (20,3% e 14,7%), respectivamente; e as estimativas das porcentagens de moradores adultos nos dois grupos de sexo com autoavaliação ruim ou muito ruim foram, respectivamente, 5,0% e 3,3%.

A partir das distribuições da Tabela 6, estima-se que a proporção de avaliação da saúde como boa ou muito boa é maior entre os homens (61,6%) do que entre as mulheres (52,4%). A autoavaliação da saúde como regular e ruim ou muito ruim é mais prevalente entre respondentes femininos (38,2% e 9,42%, respectivamente) do que respondentes masculinos (aproximadamente, 31,4% e 7%).

Tabela 5 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo o sexo e a autoavaliação da saúde, Região Nordeste, 2013.

Sexo	Autoavaliação da saúde	Proporção (%)	EP (%)	CV (%)	EPA
Feminino	Boa	27,80	0,41	1,46	1,51
	Regular	20,27	0,41	2,01	1,87
	Ruim	5,00	0,22	4,44	1,90
Masculino	Boa	28,90	0,36	1,23	1,13
	Regular	14,74	0,33	2,24	1,59
	Ruim	3,29	0,20	6,18	2,38
Total		100,00			

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Tabela 6 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional ao sexo, Região Nordeste, 2013.

Sexo	Autoavaliação da saúde	Proporção (%)	EP (%)	CV (%)	EPA
Feminino	Boa	52,38	0,77	1,46	2,48
	Regular	38,20	0,77	2,01	2,62
	Ruim	9,42	0,42	4,44	2,16
Masculino	Boa	61,58	0,76	1,23	1,89
	Regular	31,41	0,70	2,24	1,79
	Ruim	7,01	0,43	6,18	2,23

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Os resultados dos testes para a hipótese nula de independência entre autoavaliação da saúde e sexo é apresentado na Tabela 7. A incorporação da estrutura do plano amostral pela correção de Rao-Scott resultou em uma estatística de 81,5, enquanto que o valor da estatística de Pearson (não corrigida) foi de 137,9, ambas com dois graus de liberdade. Assim, o teste corrigido detecta fortes evidências estatísticas (p inferior a 1 por 10.000) para rejeitar a hipótese de independência entre a autoavaliação de saúde e o sexo dos moradores adultos no Nordeste, em 2013. Esta mesma conclusão é também obtida pelo teste F com a correção de Rao-Scott.

Tabela 7 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e o sexo na população alvo, Região Nordeste, 2013.

Estatística	Valor	Graus de liberdade	Valor p
X_P^2	159,82	2	< 0,0001
X_{RS}^2	81,45	2	< 0,0001
F_{RS}	40,73	(1,9891, 3.425,3)	< 0,0001

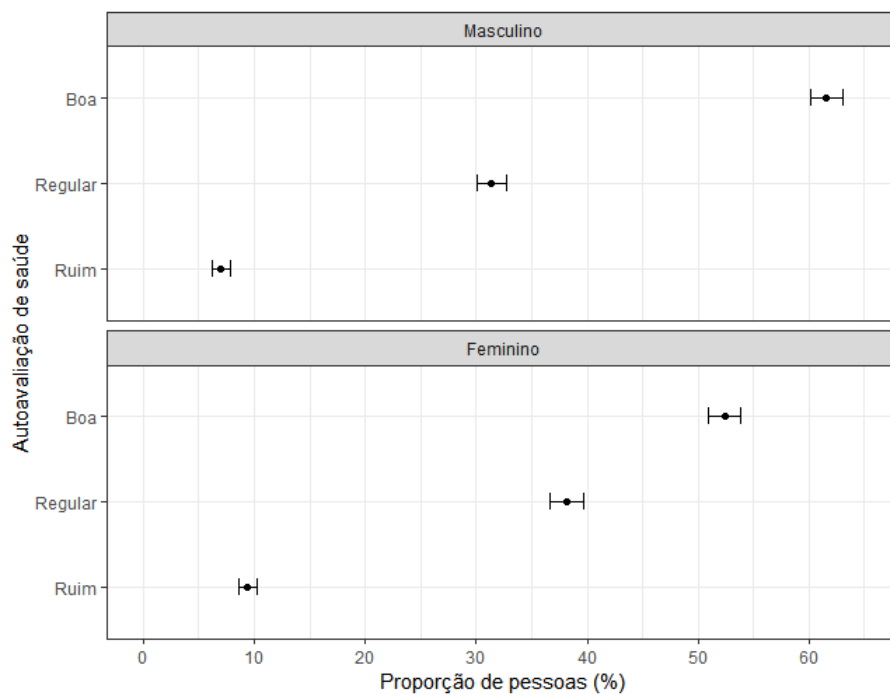
Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Para elucidar de que forma se dá a associação entre autoavaliação da saúde e sexo nesta população, foram construídos intervalos de Wald de 95% de confiança para as proporções populacionais da autoavaliação da saúde condicional aos grupos de sexo. Os correspondentes intervalos usaram o método de Wald

$$\hat{p} \pm z \text{ erro padrão}(\hat{p}),$$

em que as proporções estimadas e correspondentes erros padrões são dados na Tabela 6 e z sendo o quantil 0,975 da distribuição normal padrão. Os intervalos obtidos são mostrados na Figura 3. Assim, observa-se que a associação entre as duas variáveis ocorre pela maior proporção de pessoas com autoavaliação boa/muito boa entre homens do que entre as mulheres e a maior proporção de pessoas com autoavaliação regular ou ruim/muito ruim entre as mulheres do que entre os homens.

Figura 3 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde condicional ao sexo, Região Nordeste, 2013.



Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

3.2.2 Associação entre autoavaliação de saúde e faixa etária

Na Tabela 8 tem-se a distribuição conjunta das porcentagens estimadas, ponderadas pelos pesos amostrais, de pessoas com 18 anos ou mais segundo as faixas etárias e a saúde auto-referida. Sobressaem-se as porcentagens correspondentes aos grupos com 30 a 59 anos e saúde boa/muito boa (29,6%) e regular (20,5%, aproximadamente) e com idade 18 a 29 anos e avaliação de saúde boa ou muito boa (21,1%).

Tabela 8 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo a faixa etária e a autoavaliação da saúde, Região Nordeste, 2013.

Faixa etária	Autoavaliação de saúde	Proporção (%)	EP (%)	CV (%)	EPA
18-29 anos	Boa	21,10	0,36	1,70	1,41
	Regular	6,37	0,30	4,64	2,68
	Ruim	0,73	0,10	13,76	2,55
30-59 anos	Boa	29,60	0,50	1,68	2,17
	Regular	20,45	0,44	2,16	2,19
	Ruim	4,64	0,23	4,92	2,16
60-64 anos	Boa	2,02	0,14	6,84	1,77
	Regular	2,51	0,17	6,66	2,08
	Ruim	0,83	0,09	10,92	1,82
65-74 anos	Boa	2,53	0,17	6,66	2,11
	Regular	3,64	0,20	5,37	1,99
	Ruim	1,13	0,11	9,48	1,88
75 anos ou mais	Boa	1,45	0,12	8,28	1,84
	Regular	2,05	0,15	7,54	2,18
	Ruim	0,96	0,12	12,44	2,75
Total		100,00			

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

De acordo com a Tabela 9, em que é exibida a distribuição das proporções (%) ponderadas da autoavaliação da saúde condicional à faixa etária, percebe-se que a proporção de avaliação da saúde como boa ou muito boa foi maior entre os adultos com idade entre 18 a 29 anos (74,2%) do que nos demais grupos etários: 30 a 59 anos (54,1%), 60 a 64 anos (37,7%), 65 a 74 anos (34,6%) e idade igual ou maior que 75 anos (aproximadamente, 32,5%). A avaliação de saúde como regular e ruim/muito ruim tiveram mais respondentes nas faixas de 60 a 64 anos (46,8% e 15,4%), 65 a 74 anos (49,9% e 15,5%) e 75 anos ou mais (45,9% e 21,6%) do que nos grupos com 18 a 29 anos (22,6% e 2,6%) e 30 a 59 anos (aproximadamente, 37,4% e 8,5%).

Tabela 9 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional à faixa etária, Região Nordeste, 2013.

Faixa etária	Autoavaliação da saúde	Proporção (%)	EP (%)	CV (%)	EPA
18-29	Boa	74,82	1,05	1,41	2,62
	Regular	22,59	1,01	4,49	2,62
	Ruim	2,59	0,36	13,80	2,26
30-59	Boa	54,12	0,84	1,55	2,95
	Regular	37,39	0,80	2,13	2,84
	Ruim	8,49	0,42	4,94	2,37
60-64	Boa	37,74	2,16	5,72	2,05
	Regular	46,83	2,43	5,18	2,45
	Ruim	15,42	1,58	10,26	1,98
65-74	Boa	34,64	1,94	5,60	2,44
	Regular	49,90	2,21	4,44	2,87
	Ruim	15,46	1,39	8,97	2,15
75+	Boa	32,48	2,43	7,50	2,42
	Regular	45,95	2,74	5,97	2,72
	Ruim	21,57	2,39	11,10	3,04

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

A partir dos resultados dos testes de independência entre a autoavaliação de saúde e a faixa etária, mostrados na Tabela 10, verifica-se que a estatística do teste Qui-Quadrado com a correção por Rao-Scott (1.486,4) foi consideravelmente maior que a estatística resultante do teste de Pearson (624,18), cada uma com oito graus de liberdade, no entanto, ambas detectam fortes evidências estatísticas suficientes para rejeitar a hipótese de independência ($p < 0,0001$) entre as duas variáveis. O teste F com a correção de Rao-Scott apresenta valor da estatística próximo a 78,02 ($p < 0,0001$) e, portanto, sugere também a rejeição da hipótese de independência entre a autoavaliação de saúde e a faixa etária.

Tabela 10 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e a faixa etária na população alvo, Região Nordeste, 2013.

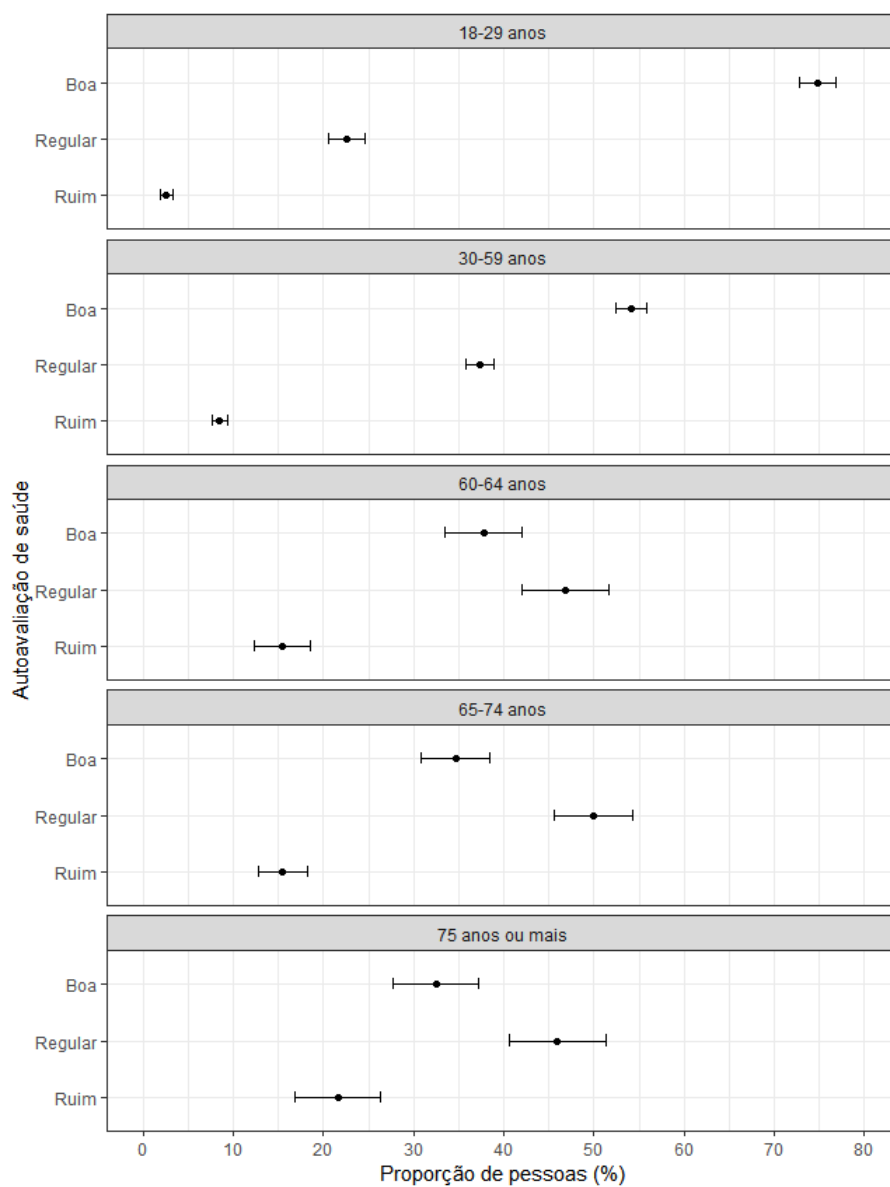
Estatística	Valor	Graus de liberdade	Valor p
X_P^2	1.486,42	8	$< 0,0001$
X_{RS}^2	624,18	8	$< 0,0001$
F_{RS}	78,02	(7,57, 13.032,0)	$< 0,0001$

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Fica evidente a associação entre a autoavaliação de saúde e os grupos de idade ao visualizar a Figura 4, em que se tem os intervalos de Wald de 95% de confiança para as proporções populacionais da autoavaliação da saúde condicional à faixa etária. Nos casos em que os intervalos não se sobrepõem, tem-se que essas proporções são significativamente diferentes. Observa-se que a proporção de adultos no Nordeste que avaliam a

sua saúde como muito boa ou boa é maior na faixa de 18 a 29 anos em comparação com as demais faixas etárias. Nota-se ainda uma tendência de maior proporção de pessoas com autoavaliação ruim ou muito ruim a medida que a idade aumenta.

Figura 4 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde condicional à faixa etária, Região Nordeste, 2013.



Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

3.2.3 Associação entre a autoavaliação de saúde e raça

Conforme a distribuição simultânea das porcentagens estimadas, ponderadas pelos pesos amostrais, segundo a raça e a saúde auto-referida, verifica-se, na Tabela 11, que 33,7% dos respondentes são de raça parda com saúde avaliada como muito boa ou boa, aproximadamente 21% são pardos e avaliam a sua saúde como regular, 15,8% são brancos e tem a sua saúde como muito boa ou boa e os demais 29,5% dos nordestinos se distribuem entre as raças branca, preta, parda e amarela/indígena com avaliação de saúde muito boa/boa, regular e ruim/muito ruim.

Tabela 11 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo a raça e a autoavaliação da saúde, Região Nordeste, 2013.

Raça	Autoavaliação de saúde	Proporção (%)	EP (%)	CV (%)	EPA
Branca	Boa	15,78	0,45	2,84	2,76
	Regular	9,15	0,35	3,77	2,63
	Ruim	2,01	0,14	7,03	1,86
Preta	Boa	6,38	0,28	4,47	2,49
	Regular	4,39	0,30	6,78	3,87
	Ruim	1,27	0,14	10,79	2,73
Amarela ou Indígena	Boa	0,81	0,11	13,58	2,75
	Regular	0,52	0,08	15,60	2,34
	Ruim	0,13	0,03	22,70	1,24
Parda	Boa	33,73	0,57	1,70	2,70
	Regular	20,95	0,54	2,56	3,17
	Ruim	4,88	0,25	5,18	2,52
Total		100,00			

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

A distribuição das porcentagens estimadas de moradores adultos da autoavaliação de saúde condicional à faixa etária são dadas na Tabela 12. A avaliação de saúde como muito boa ou boa teve as seguintes proporções por raça: amarela ou indígena (55,2%), branca (58,6%), parda (56,6%) e preta (53%). Estima-se que, no ano 2013, a avaliação de saúde como regular no nordeste brasileiro foi maior entre a raça preta (36,5%), seguida das raças amarela/indígena (35,8%), parda (35,2%) e branca (34%). A proporção de avaliação da saúde como ruim ou muito ruim não teve nítida diferença entre as raças, apresentando os valores: preta (10,5%), amarela ou indígena (9%), parda (8,2%) e branca (7,5%).

Tabela 12 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional à raça, Região Nordeste, 2013.

Raça	Autoavaliação da saúde	Proporção (%)	EP (%)	CV (%)	EPA
Amarela ou Indígena	Boa	55,22	4,71	8,52	2,40
	Regular	35,80	4,61	12,87	2,48
Branca	Ruim	8,98	2,02	22,54	1,34
	Boa	58,57	0,92	1,58	1,79
	Regular	33,96	0,91	2,68	1,88
Parda	Ruim	7,48	0,52	6,97	2,00
	Boa	56,64	0,75	1,32	2,48
	Regular	35,17	0,73	2,06	2,50
Preta	Ruim	8,19	0,41	4,97	2,39
	Boa	52,98	1,64	3,10	2,30
	Regular	36,50	1,66	4,56	2,53
	Ruim	10,52	1,08	10,30	2,64

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

De acordo com a Tabela 13, tem-se que a estatística do teste de Pearson é igual a 33,5 com seis graus de liberdade ($p < 0,0001$), indicando fortes evidências para se rejeitar a hipótese de independência entre a autoavaliação de saúde e a raça dos adultos no nordeste em 2013. Contudo, os testes considerando a estrutura do plano amostral, produz o valor da estatística de Pearson corrigida igual a 13,6 (redução de aproximadamente 58% do valor de X_P^2) com seis graus de liberdade e o da estatística F corrigida igual a 2,26 com graus de liberdade 5,6 e 9789,3 aproximadamente. Os valores p destas estatísticas são 0,0350 e 0,0382, respectivamente, sugerindo evidências estatísticas apenas moderadas para rejeitar a hipótese nula de independência.

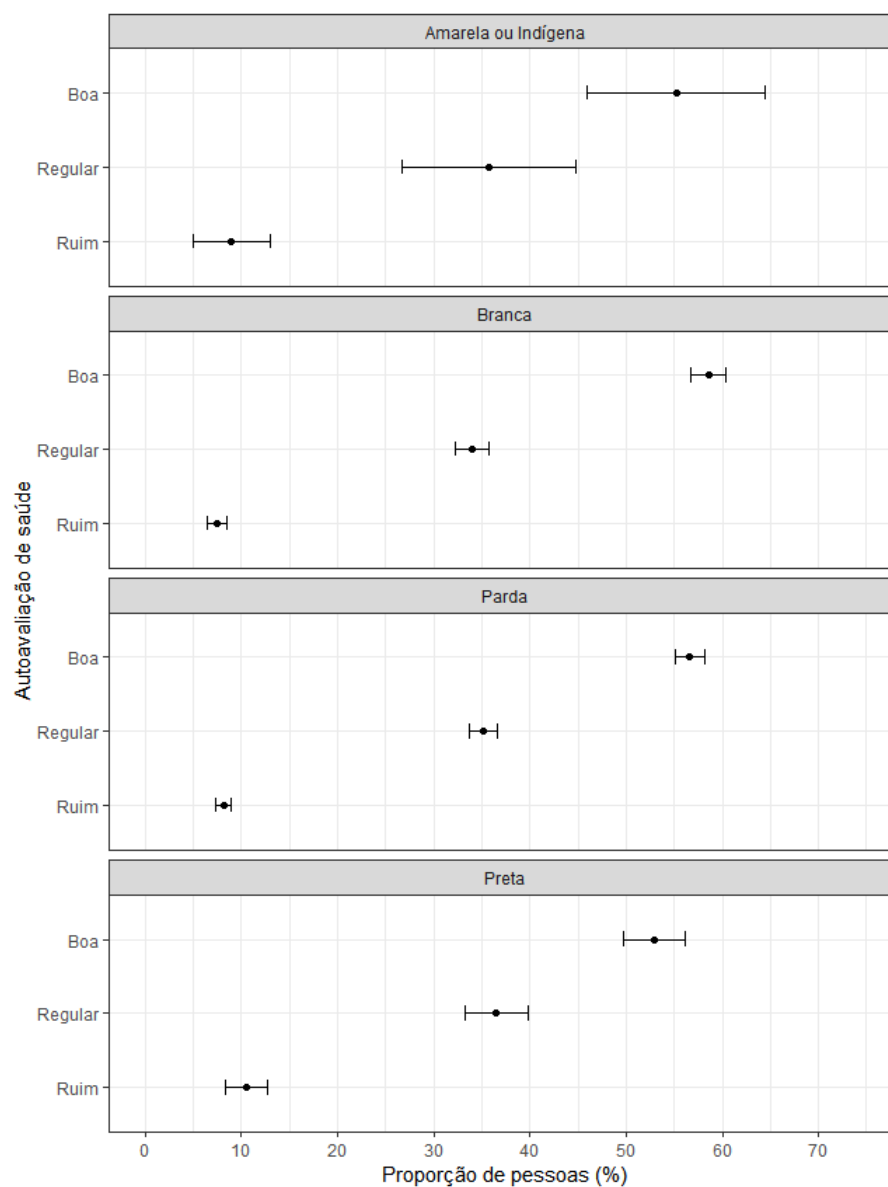
Tabela 13 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e raça na população alvo, Região Nordeste, 2013.

Estatística	Valor	Graus de liberdade	Valor p
X_P^2	29,08	6	$< 0,0001$
X_{RS}^2	13,55	6	0,0350
F_{RS}	2,26	(5,6848, 9.789,3)	0,0382

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

É possível perceber, a partir dos intervalos de 95% de confiança dados na Figura 5, que as distribuições das proporções ponderadas da autoavaliação de saúde condicional aos grupos de raça são aproximadamente homogêneas, sugerindo que não há evidências suficientes para declarar que estas duas variáveis estão associadas.

Figura 5 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde e raça, Região Nordeste, 2013.



Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

3.2.4 Associação entre autoavaliação de saúde e nível de instrução

A distribuição conjunta das porcentagens estimadas dos moradores adultos no nordeste segundo o nível de instrução e a autoavaliação da saúde é dada na Tabela 14. Verifica-se que as maiores porcentagens estimadas ocorrem nas categorias: sem escolaridade/ensino fundamental incompleto e saúde boa/muito boa (21,1%) e regular (20,5%) e médio completo/superior incompleto e saúde muito boa/boa (20,2%).

Tabela 14 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo o nível de instrução e a autoavaliação da saúde, Região Nordeste, 2013.

Nível de instrução	Autoavaliação da saúde	Proporção (%)	EP (%)	CV (%)	EPA
Sem instrução ou fundamental incompleto	Boa	21,07	0,50	2,38	2,76
	Regular	20,48	0,55	2,69	3,41
	Ruim	6,53	0,28	4,28	2,34
Fundamental completo ou médio incompleto	Boa	9,35	0,37	3,94	2,94
	Regular	4,47	0,29	6,56	3,69
	Ruim	0,54	0,07	12,04	1,45
Médio completo ou superior incompleto	Boa	20,22	0,49	2,45	2,78
	Regular	8,09	0,34	4,22	2,87
	Ruim	0,99	0,11	11,15	2,27
Superior completo	Boa	6,06	0,29	4,71	2,62
	Regular	1,98	0,16	8,19	2,48
	Ruim	0,23	0,05	23,12	2,21
Total		100,00			

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

De acordo com as porcentagens estimadas da autoavaliação de saúde condicional ao nível de escolaridade, dadas na Tabela 15, estima-se que as proporções estimadas de avaliação da saúde boa ou muito boa são respectivamente 43,8%, 65,1%, 69,0% e 73,3% em relação as quatro categorias da escolaridade, indicando que a autoavaliação positiva da saúde aumenta com a escolaridade.

Tabela 15 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional ao nível de instrução, Região Nordeste, 2013.

Nível de instrução	Autoavaliação da saúde	Proporção (%)	EP (%)	CV (%)	EPA
Sem instrução ou fundamental incompleto	Boa	43,82	0,89	2,03	2,73
	Regular	42,59	0,93	2,18	3,01
	Ruim	13,59	0,56	4,13	2,29
Fundamental completo ou médio incompleto	Boa	65,10	1,75	2,69	3,54
	Regular	31,12	1,71	5,50	3,59
	Ruim	3,78	0,45	11,81	1,44
Médio completo ou superior incompleto	Boa	69,02	1,07	1,55	2,88
	Regular	27,61	1,02	3,70	2,83
	Ruim	3,37	0,36	10,78	2,20
Superior completo	Boa	73,34	1,66	2,26	2,45
	Regular	23,93	1,59	6,64	2,42
	Ruim	2,73	0,62	22,85	2,55

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Na Tabela 16, são apresentados os resultados dos testes para a hipótese nula de independência entre autoavaliação da saúde e nível de instrução. Os testes pelas estatísticas de Pearson sem correção e corrigida pelo método de Rao-Scott produzem diferentes valores das estatísticas (respectivamente 1.373,1 e 552,47, ambas com seis graus de liberdade). O valor da estatística de teste F corrigido é igual a 92,1. Todos esses testes fornecem fortes evidências estatísticas ($p < 0,0001$) para rejeitar a hipótese de independência entre a autoavaliação de saúde e o nível de instrução dos moradores adultos no Nordeste, em 2013.

Tabela 16 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e o nível de instrução na população alvo, Região Nordeste, 2013.

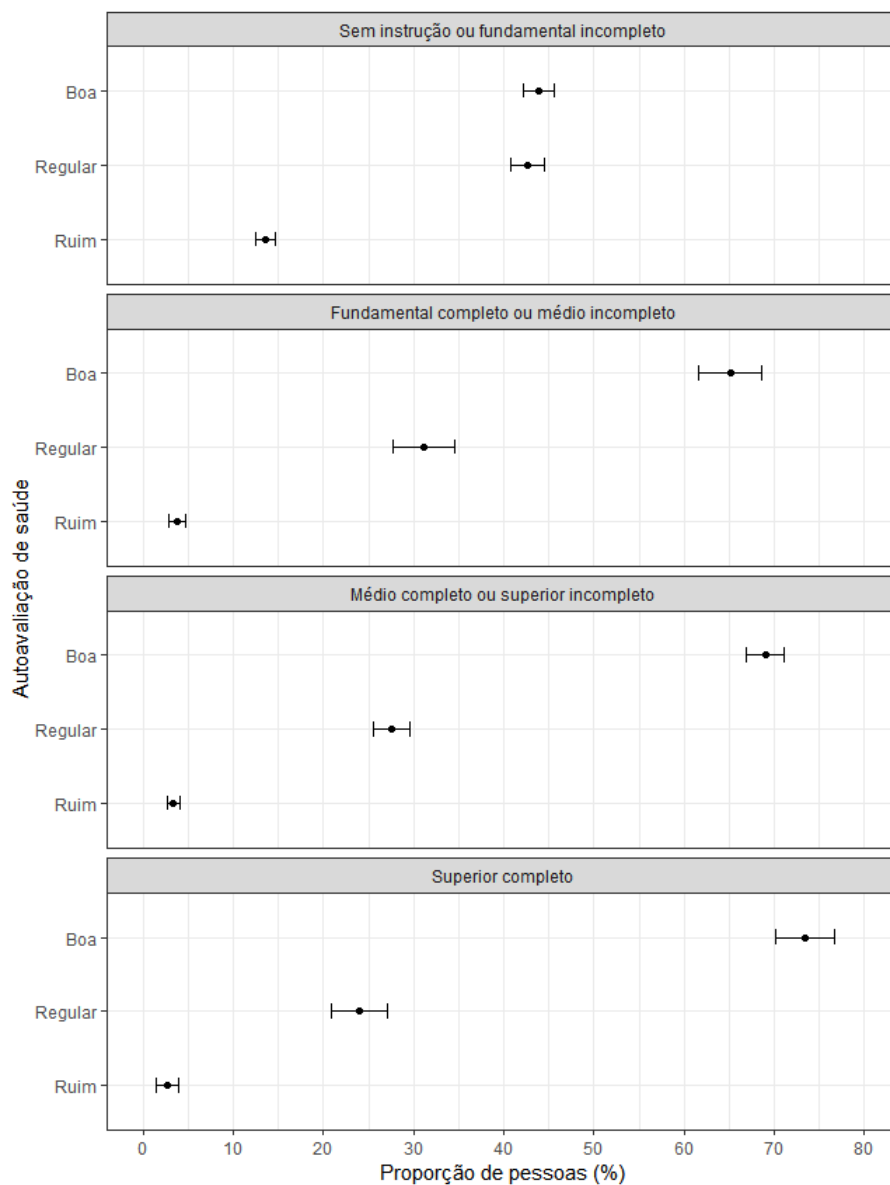
Estatística	Valor	Graus de liberdade	Valor p
X_P^2	1.375,48	6	$< 0,0001$
X_{RS}^2	552,47	6	$< 0,0001$
F_{RS}	92,08	(5,3617, 9.232,9)	$< 0,0001$

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Analisando-se os intervalos de Wald de 95% de confiança para as proporções populacionais da saúde auto-referida condicional às categorias do nível de instrução, dados na Figura 6, observa-se que a associação entre as variáveis pode ser explicada pela menor proporção de avaliação muito boa ou boa entre os moradores sem instrução ou ensino fundamental incompleto do que as correspondentes proporções entre aqueles moradores com maior escolaridade. Também, observa-se as proporções de avaliação regular ou ruim/muito

ruim entre os moradores sem instrução ou ensino fundamental incompleto são significativamente superior as respectivas proporções para os indivíduos de maior escolaridade.

Figura 6 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde e nível de instrução, Região Nordeste, 2013.



Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

3.2.5 Associação entre autoavaliação de saúde e situação habitacional

A distribuição das porcentagens ponderadas, estimadas conjuntamente segundo a situação habitacional e a autoavaliação da saúde, é exibida na Tabela 17. Nota-se que as porcentagens mais elevadas ocorrem na situação do tipo urbana e avaliação de saúde como muito boa/boa (44,8%) e regular (25,7%); seguidas das porcentagens referentes à situação rural e avaliação de saúde muito boa/boa (11,9%) e regular (9,3%). As porcentagens de adultos nas situações habitacionais urbana e rural com autoavaliação ruim/muito ruim foram, respectivamente, 5,8% e 2,5%.

Tabela 17 – Proporção estimada de pessoas de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) da proporção estimada, segundo a situação habitacional e a autoavaliação da saúde, Região Nordeste, 2013.

Situação	Autoavaliação da saúde	Proporção (%)	EP (%)	CV (%)	EPA
Rural	Boa	11,89	0,40	3,35	2,77
	Regular	9,34	0,33	3,50	2,31
	Ruim	2,53	0,17	6,81	2,21
Urbano	Boa	44,81	0,58	1,28	2,46
	Regular	25,67	0,54	2,10	2,79
	Ruim	5,76	0,27	4,65	2,42
Total		100,00			

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Na Tabela 18 são apresentadas as porcentagens estimadas, com a correção dos pesos amostrais, da autoavaliação da saúde condicional à situação habitacional, para o nordeste do Brasil no ano 2013. Verifica-se que a proporção de avaliação de saúde como boa ou muito boa foi maior entre a situação urbana (58,8%) do que entre a rural (50%). A avaliação como regular e ruim/muito ruim teve maior proporção entre os indivíduos que moram na zona rural (39,3% e 10,7%, respectivamente) do que os moradores na zona urbana (33,7% e 7,6%).

Tabela 18 – Proporção estimada de moradores de 18 anos ou mais, erro padrão (EP), coeficiente de variação (CV) e efeito do plano amostral (EPA) para as categorias da autoavaliação de saúde condicional à situação habitacional, Região Nordeste, 2013.

Situação habitacional	Autoavaliação da saúde	Proporção (%)	EP (%)	CV (%)	EPA
Rural	Boa	50,04	1,13	2,26	2,16
	Regular	39,31	1,14	2,89	2,29
	Ruim	10,65	0,72	6,78	2,32
Urbano	Boa	58,77	0,67	1,14	2,61
	Regular	33,67	0,65	1,94	2,70
	Ruim	7,56	0,35	4,61	2,45

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

De acordo com a Tabela 19, as estatísticas de Pearson (147,32; 2 g.l.), Pearson ajustada por Rao-Scott (47,03; 2 g.l.) e F corrigido por Rao-Scott (23,51) encontram fortes evidências ($p < 0,0001$) para a rejeição da hipótese nula de independência entre a autoavaliação de saúde e a situação habitacional na região Nordeste no ano 2013.

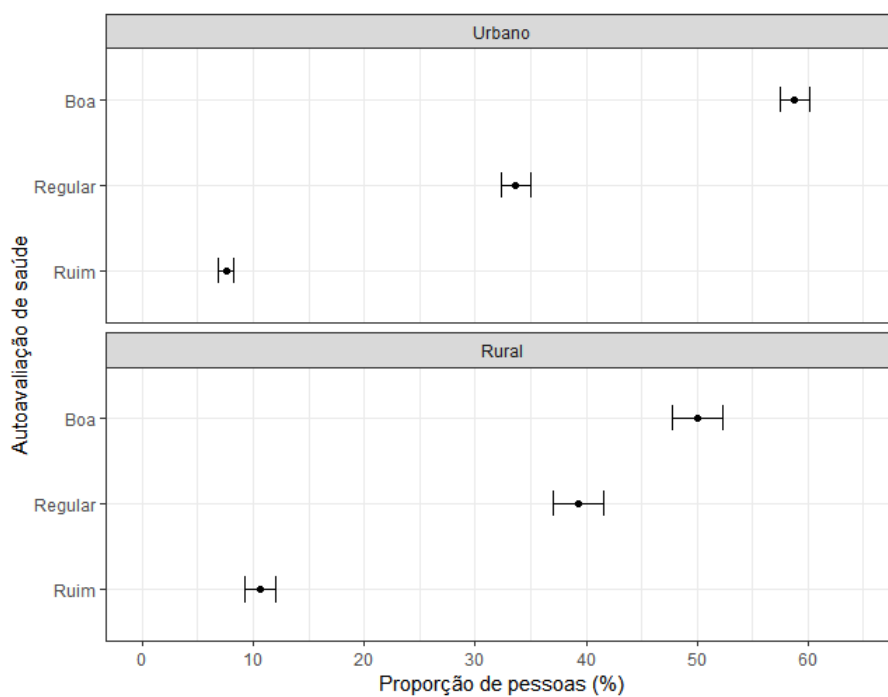
Tabela 19 – Estatísticas de Pearson não corrigida e estatísticas de Pearson e F com correção de Rao-Scott para o teste da hipótese de independência entre autoavaliação de saúde e situação habitacional na população alvo, Região Nordeste, 2013.

Estatística	Valor	Graus de liberdade	Valor p
X_P^2	112,87	2	$< 0,0001$
X_{RS}^2	47,03	2	$< 0,0001$
F_{RS}	23,51	(1,995, 3.435,1)	$< 0,0001$

Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

Com o auxílio dos intervalos de Wald para as proporções populacionais estimadas, dados na Figura 7, observa-se que a associação entre as duas variáveis ocorre pelo fato de moradores que residem em domicílios situados na zona urbana têm maior proporção de pessoas com autoavaliação boa/muito boa e menores proporções de autoavaliação regular e ruim/muito ruim do que os moradores da zona rural.

Figura 7 – Intervalos de 95% de confiança para as proporções estimadas da distribuição da autoavaliação do estado de saúde e situação habitacional, Região Nordeste, 2013.



Fonte: IBGE, Pesquisa Nacional de Saúde, 2013.

4 Considerações finais

Dados de acesso público são facilmente encontrados hoje em dia sobre vários aspectos da sociedade. Um desafio importante para algum usuário destes dados é a realização de análises adequadas desses dados que permitam a retirada de conclusões válidas. Tais análises são geralmente possíveis quando se incorpora as características do planejamento e coleta dos dados coletados. Uma das bases de dados que se insere neste contexto é a Pesquisa Nacional de Saúde (PNS), que é disponibilizada pelo IBGE e oferece muitas informações sobre a saúde da população brasileira.

Neste trabalho, utilizando o *software R*, foi possível ler e analisar os dados oriundos da PNS, realizada em 2013 com base em uma amostra estratificada com seleção de conglomerados em múltiplos estágios. Uma das características dessa pesquisa é a necessidade de inclusão de fatores expansão variáveis para estimação de quantidades populacionais. Estes fatores, além de compensar pela parte da população não amostrada, são baseados em ajustes para tratar não-resposta e realizar calibrações.

Dentre as características medidas pela PNS, destaca-se a medida de autoavaliação da saúde, em que os moradores respondentes informam a percepção sobre o estado de saúde deles. Usando apenas dados da região Nordeste da população alvo da PNS, o trabalho teve como objetivos testar se autoavaliação independe dos fatores socio-demográficos: sexo, idade, raça, escolaridade e situação habitacional do domicílio e, nos casos de associação detectada, descrever como se dá a associação entre as variáveis. O processo de estimação para responder estes objetivos comparou estimadores e testes como se os dados tivessem sido coletados através de uma amostra aleatória simples com os estimadores e testes apropriados para os dados da PNS, os quais incorporam a estrutura do plano amostral e pesos amostrais da da pesquisa.

Com base nos resultados obtidos, pode-se observar que a autoavaliação de saúde e sexo dos moradores adultos no nordeste estão fortemente associados, em que mais homens do que mulheres declaram a saúde como muito boa ou boa ao passo que mais mulheres do que homens avaliam a sua saúde como regular ou ruim/muito ruim.

A autoavaliação de saúde está significativamente associada com a faixa etária. Foi observado que, com o aumento da idade, maior é a proporção de residentes com saúde referida como ruim ou muito ruim, como também se verificou uma declínio das avaliações como positivas com o aumento da idade.

Em relação a raça, os testes com o ajuste de Rao-Scott detectaram evidências estatísticas moderadas para rejeitar a independência entre a autoavaliação de saúde e a raça.

A independência entre a autoavaliação de saúde e a escolaridade foi fortemente re-

jeitada. A proporção de pessoas que avaliam a sua saúde como ruim ou muito ruim foi significativa entre pessoas que não possuem instrução ou ainda não concluíram o ensino fundamental do que entre os outros níveis de escolaridade. Ainda se observou que quanto maior o nível de escolaridade, maior a proporção de avaliação do estado de saúde como bom ou muito bom.

A situação habitacional apresentou associação significativa com a saúde auto-referida, em que a proporção de pessoas com avaliação muito boa ou boa foi maior entre os que vivem na zona urbana enquanto que a proporção de avaliação da saúde como regular ou ruim/muito ruim foi maior entre os residentes da zona rural.

Todas estas interpretações sugerem a possibilidade de associações entre as variáveis investigadas neste estudo. Contudo, em virtude da PNS ser um estudo observacional, tais associações podem não implicar relações de causa-e-efeito.

Para trabalhos futuros, pode-se procurar estender as análises deste trabalho para outras regiões do país, bem como avaliar se as associações entre as variáveis estudadas se mantêm dentre dos estados de cada região. Também, análises alternativas podem ser realizadas com a aplicação de modelos de regressão logística e modelos log-lineares.

Referências

- BARROS, M. B. A. et al. Auto-avaliação da saúde e fatores associados, Brasil, 2006. *Revista de Saúde Pública*, v. 43, p. 27–37, 2009.
- BURSTRÖM, B.; FREDLUND, P. Self rated health: Is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes? *Journal of Epidemiology & Community Health*, v. 55, n. 11, p. 836–840, 2001.
- DURBIN, J. Some results in sampling theory when the units are selected with unequal probabilities. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 15, n. 2, p. 262–269, 1953.
- FREITAS, M. P. S. Pesquisa nacional de saúde: Plano amostral. Diretoria de Pesquisas, Coordenação de Métodos e Qualidade, Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, 2014.
- HANSEN, M. H.; HURWITZ, W. N.; MADOW, W. G. *Sample survey methods and theory. Volume 1: Methods and applications*. New York: Wiley, 1953.
- _____. *Sample survey methods and theory. Volume 2: Theory*. New York: Wiley, 1953.
- HEERINGA, S.; WEST, B.; BERGLUND, P. *Applied Survey Data Analysis*. New York: Chapman and Hall/CRC, 2010.
- IBGE. *Pesquisa Nacional de Saúde 2013: Percepção do estado de saúde, estilos de vida e doenças crônicas. Brasil, grandes regiões, unidades da federação, municípios das capitais e distrito federal*. Rio de Janeiro, 2014. 181 p.
- IDLER, E. L.; BENYAMINI, Y. Self-rated health and mortality: a review of twenty-seven community studies. *Journal of health and social behavior*, p. 21–37, 1997.
- LEHTONEN, R.; PAHKINEN, E. *Practical Methods for Design and Analysis of Complex Surveys*. Second edition. Chichester: Wiley, 2004.
- LOHR, S. L. *Sampling: Design and Analysis*. Second edition. Boston, MA: Brooks/Cole, Cengage Learning, 2010.
- LUMLEY, T. *Complex surveys: a guide to analysis using R*. Hoboken: Wiley, 2011.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018.
- SOUZA-JÚNIOR, P. R. B. et al. Desenho da amostra da pesquisa nacional de saúde 2013. *Epidemiologia e Serviços de Saúde*, v. 24, p. 207–216, 2015.
- SZWARCWALD, C. L. et al. Pesquisa nacional de saúde no Brasil: concepção e metodologia de aplicação. *Ciênc. saúde coletiva*, v. 19, p. 333 – 342, 02 2014.

THEME-FILHA, M. M.; SZWARCOWALD, C. L.; SOUZA-JÚNIOR, P. R. B. Socio-demographic characteristics, treatment coverage, and self-rated health of individuals who reported six chronic diseases in Brazil, 2003. *Cadernos de Saúde Pública*, v. 21, p. S43 – S53, 2005.

_____. Medidas de morbidade referida e inter-relações com dimensões de saúde. *Revista de Saúde Pública*, v. 42, p. 73 – 81, 2008.

WOLTER, K. M. *Introduction to Variance Estimation*. Second edition. New York: Springer, 2007.

Apêndices

Apêndice A - Variáveis

- v0001 (Unidade de Federação)
 - 21 - Maranhão
 - 22 - Piauí
 - 23 - Ceará
 - 24 - Rio Grande do Norte
 - 25 - Paraíba
 - 26 - Pernambuco
 - 27 - Alagoas
 - 28 - Sergipe
 - 29 - Bahia
- v0024 (Identificador do estrato)
- upa_pns (Identificador da UPA)
- v0025 (Seleção do morador de 18 anos ou mais para responder o questionário individual)
 - 1 - Selecionado
 - 0 - Não selecionado
- v0026 (Tipo de situação censitária do domicílio)
 - 1 - Urbano
 - 2 - Rural
- c006 (Sexo do morador)
 - 1 - Masculino
 - 2 - Feminino
- c008 (Idade em anos do morador)
- c009 (Cor ou raça do morador)
 - 1 - Branca
 - 2 - Preta

- 3 - Amarela
- 4 - Parda
- 5 - Indígena
- 9 - Ignorado
- vdd004 (Nível de instrução do morador)
 - 1 - Sem instrução
 - 2 - Fundamental incompleto ou equivalente
 - 3 - Fundamental completo ou equivalente
 - 4 - Médio incompleto ou equivalente
 - 5 - Médio completo ou equivalente
 - 6 - Superior completo ou equivalente
 - 7 - Superior completo
- n001 (Em geral, como o(a) sr(a) avalia a sua saúde?)
 - 1 - Muito boa
 - 2 - Boa
 - 3 - Regular
 - 4 - Ruim
 - 5 - Muito ruim
- v0028 (Peso do domicílio com correção de não entrevista sem calibração pela projeção de população)
- v0029 (Peso do morador selecionado com correção de não entrevista sem calibração pela projeção de população para morador selecionado)
- v00282 (Projeção da população)
- v00283 (Domínio de pós-estrato 1)
- v00292 (Projeção da população obtida com os dados dos moradores dos domicílios da pesquisa - só aplicável para domicílios com entrevista realizada para morador selecionado)
- v00293 (Domínio de pós-estrato 2 - só aplicável para domicílios com entrevista realizada para morador selecionado)

Apêndice B - Código R

```
#-----#
### Programa R para reproduzir as análises do Capítulo 3 do trabalho:
### Nascimento, D. L. (2018) "Autoavaliação da Saúde no Nordeste
### Brasileiro e Fatores Sociodemográficos: Um Estudo de Associação
### com base na Pesquisa Nacional de Saúde, 2013".
### Monografia de graduação. Departamento de Estatística.
### Universidade Federal do Rio Grande do Norte.
###
### Orientador: Prof. Dr. Damião Nóbrega da Silva
#-----#

##### Definição do diretório #####

setwd("F:/tcc/")

##### Instalação de pacotes #####
install.packages( c( "downloader" , "SAScii" , "dplyr" , "survey" ,
                    "xtable" , "ggplot2" , "plotrix" ) )

##### Carregamento de pacotes #####
library(downloader)
library(SAScii)
library(dplyr)
library(survey)
library(xtable)
library(ggplot2)
library(plotrix)

##### Comandos adaptados de Djalma Pessoa e Anthony J Damico através
##### do link: http://asdfree.com/pesquisa-nacional-de-saude-pns.html
##### Download e leitura dos arquivos e banco de dados #####
source(
  "https://raw.githubusercontent.com/ajdamico/asdfree/archive/Download%20Cache/
  download%20cache.R" ,
  prompt = FALSE ,
  echo = FALSE
)
```

```

# download dos microdados da PNS
tf <- "F:/tcc.zip"

download_cached(
  "ftp://ftp.ibge.gov.br/PNS/2013/microdados/pns_2013_microdados_2017_03_23.zip",
  tf , mode = 'wb' )

z <- unzip( tf , exdir = "F:/tcc" )

# leitura de dados
dd <- "./Dados/DOMPNS2013.txt"
ds <- "./Dicionarios_e_input/input_DOMPNS2013.txt"
pd <- "./Dados/PESPNS2013.txt"
ps <- "./Dicionarios_e_input/input_PESPNS2013.txt"

# cria um objeto do tipo data.frame 'dom' contendo um registro por domicílio
dom <- read.SAScii( dd , ds )

# cria um objeto do tipo data.frame 'pes' contendo um registro por pessoa
pes <- read.SAScii( pd , ps )

# converte os nomes das colunas para letra minúscula
names( dom ) <- tolower( names( dom ) )
names( pes ) <- tolower( names( pes ) )

# pessoas - peso pré-estratificado
names( pes )[ names( pes ) == 'v0029' ] <- 'pre_pes_long'
names( pes )[ names( pes ) == 'v0028' ] <- 'pre_pes_full'

# une os arquivos dom e pes
x <- merge( dom , pes , by = c( "v0001" , "v0024" , "upa_pns" , "v0006_pns" ) )

stopifnot( nrow( x ) == nrow( pes ) )

##### Preparação das variáveis #####
# região
x <- transform( x , region = factor( substr( v0001 , 1 , 1 ) ,
  labels = c( "Norte" , "Nordeste" , "Sudeste" , "Sul" , "Centro-Oeste" ) ) )

# Unidade da Federação
x <- transform( x , uf = factor( v0001 , labels = c(

```

```

"Rondônia", "Acre", "Amazonas", "Roraima", "Pará", "Amapá", "Tocantins",
"Maranhão", "Piauí", "Ceará", "Rio Grande do Norte", "Paraíba",
"Pernambuco", "Alagoas", "Sergipe", "Bahia",
"Minas Gerais", "Espírito Santo", "Rio de Janeiro", "São Paulo",
"Paraná", "Santa Catarina", "Rio Grande do Sul", "Mato Grosso do Sul",
"Mato Grosso", "Goiás", "Distrito Federal") ) )

# Estrato - situação habitacional
x <- transform( x, situ = ifelse( v0026 == "1", "Urbano", "Rural") )

# Autoavaliação do estado de saúde
x <- transform( x, saude = ifelse( n001 == 1 | n001 == 2 , "Boa",
  ifelse( n001==3, "Regular",
  ifelse ( n001 == 4 | n001 == 5, "Ruim", "NA") ) ) )

# Sexo
x <- transform( x, sexo = ifelse( c006 == "1" , "Masculino", "Feminino") )

# Raça
x <- transform( x, raca =
  ifelse( c009 == 1, "Branca",
  ifelse( c009 == 2, "Preta",
  ifelse( c009 == 4, "Parda",
  ifelse( c009 == 3 | c009 == 5, "Amarela ou Indígena", "NA") ) ) ) )

# Nível de instrução
x <- transform( x, instrucao =
  ifelse( vdd004 == 1 | vdd004== 2,
  "Sem instrução ou Fundamental incompleto",
  ifelse( vdd004 == 3 | vdd004 == 4,
  "Fundamental completo ou médio incompleto",
  ifelse( vdd004 == 5 | vdd004 == 6,
  "Médio completo ou superior incompleto",
  ifelse( vdd004 == 7, "Superior completo", "NA") ) ) ) )

x$c008 <- as.numeric(x$c008)

# coluna com valor um
x$one <- 1

##### Características básicas da PNS #####

```

```
#Nacional
# Quantidade de UPA's - país
length( table( x$upa_pns ))

# Quantidade de estratos - país
length( table( x$v0024))

# quantidade de domicílios com morador selecionado no país
table( x$v0025 == 1)

# Quantidade de entrevista individual realizada no país
table( x$m001 == 1)

# Região Nordeste

nordeste <- subset( x, region == "Nordeste", na.rm=T)

# Quantidade de UPA's - região Nordeste
length( table( nordeste$upa_pns ))

# Quantidade de estratos - região Nordeste
length( table( nordeste$v0024))

# quantidade de domicílios com morador selecionado no Nordeste
table( nordeste$v0025 == 1)

# Quantidade de entrevista individual realizada no Nordeste
table( nordeste$m001 == 1)

##### Plano com os respondentes da região Nordeste #####

nordeste <- subset( nordeste, m001 == "1", na.rm=T)

# Faixa etária
nordeste <- transform( nordeste, idade = factor(
  1 + findInterval( c008, c( 18, 30, 60, 65, 75)),
  labels = c("18-29", "30-59", "60-64", "65-74", "75+"))

# plano amostral pré-estratificado
```

```

plano_pre <-
  svydesign(
    id = ~ upa_pns ,
    strata = ~ v0024 ,
    data = nordeste ,
    weights = ~ pre_pes_long ,
    nest = TRUE
  )

# figure out stratification targets
p_pop <- unique( nordeste [ c( 'v00293.y' , 'v00292.y' ) ] )

names( p_pop ) <- c( "v00293.y" , "Freq" )

# plano amostral pós-estratificado
plano_no <- postStratify( plano_pre , ~v00293.y , p_pop , na.rm=T)

##### Regiao Nordeste #####

# Histograma para a variável idade sem e com os pesos amostrais
par( mfrow = c(1,2) )
hist( nordeste$c008, xlab = "Idade (anos)", freq = F,
      ylab = "Densidade", main = "", ylim = c(0,0.03), density=30 )
axis.break( axis = 1, breakpos = 18, style = "zigzag" )
svyhist( ~c008, plano_no, ylab = "Densidade", freq = F, density=30,
         main = "", xlab = "Idade (anos)", ylim = c(0,0.03) )
axis.break( axis = 1, breakpos = 18, style = "zigzag" )

##### Estatísticas descritivas para os pesos amostrais #####
summary( weights(plano_no) )
# Percentis
quantile( weights(plano_no) , c(.01, .05, .25, .5, .75, .95, .99))
# soma dos pesos amostrais
sum( weights(plano_no) )
# tamanho da amostra - respondentes
sum( weights(plano_no, "sampling") !=0 )

# Boxplot dos pesos amostrais por UF - Região Nordeste, PNS 2013

par( mfrow = c(1,1) )

```

```

svyboxplot( weights(plano_no) ~ as.character(uf), plano_no, axes=F,
  col="gray", xlab = "Unidades da federação", ylab="Peso amostral")
axis(1, 1:9, labels=c("AL", "BA", "CE", "MA", "PB", "PE", "PI", "RN", "SE"))
axis(2); box()

##### Análise exploratória univariada #####

n_amostra <- c(coef(svyby( ~ one , ~ sexo, plano_no, unwtd.count)),
  coef(svyby( ~ one , ~ idade, plano_no, unwtd.count)),
  coef(svyby( ~ one , ~ raca, plano_no, unwtd.count)),
  coef(svyby( ~ one , ~ instrucao, plano_no, unwtd.count)),
  coef(svyby( ~ one , ~ situ, plano_no, unwtd.count)),
  coef(svyby( ~ one , ~ saude, plano_no, unwtd.count)))

p_amostral <- c(
  table(nordeste$sexo)/sum( table(nordeste$sexo) ) * 100,
  table(nordeste$idade)/sum( table(nordeste$idade) ) * 100,
  table(nordeste$raca)/sum( table(nordeste$raca) ) * 100,
  table(nordeste$instrucao)/sum(table(nordeste$instrucao))*100,
  table(nordeste$situ)/sum( table(nordeste$situ) ) * 100,
  table(nordeste$saude)/sum( table(nordeste$saude) ) * 100 )

p_com_pesos <- c( coef(svymean ( ~ sexo , plano_no, deff=T)) * 100,
  coef(svymean ( ~ idade , plano_no, deff=T)) * 100,
  coef(svymean ( ~ raca , plano_no, deff=T)) * 100,
  coef(svymean ( ~ instrucao , plano_no, deff=T)) * 100,
  coef(svymean ( ~ situ , plano_no, deff=T)) * 100,
  coef(svymean ( ~ saude , plano_no, deff=T)) * 100)

coef_var <- c( cv( svymean( ~ sexo , plano_no)) * 100,
  cv( svymean( ~ idade , plano_no)) * 100,
  cv( svymean( ~ raca , plano_no)) * 100,
  cv( svymean( ~ instrucao , plano_no)) * 100,
  cv( svymean( ~ situ , plano_no)) * 100,
  cv( svymean( ~ saude , plano_no)) * 100)

epa <- c(0.0, 0.0, 0.6917, 0.5653, 1.6229, 1.5058, 1.7529, 2.4489,
  3.8637, 4.3957, 3.9440, 2.9479, 2.9500,
  3.2121, 3.0277, 2.6403, 2.6403, 2.4510, 2.5687, 2.3658)
tab_univariada <- data.frame(n_amostra, p_amostral, p_com_pesos, coef_var,epa)
options(OutDec=",")

```

```
xtable( tab_univariada , digits = 1 )
options(OutDec=".")

##### Análise exploratória bivariada #####

# Distribuição da autoavaliação da saúde | sexo
saude_sexo2 <- svyby( ~saude , ~sexo, plano_no, svymean, deff = T)
p_saude_sexo2 <- coef(saude_sexo2)
ep_saude_sexo2 <- c( SE(saude_sexo2)[,1], SE(saude_sexo2)[,2],
                    SE(saude_sexo2)[,3])
coef_sexo2 <- c( cv(saude_sexo2)[,1], cv(saude_sexo2)[,2], cv(saude_sexo2)[,3])
epa_saude_sexo2 <- c(2.484007, 1.886960, 2.623768, 1.789267,
2.158844, 2.234960)

tab_saude_sexo2 <- data.frame( p_saude_sexo2 * 100,
                              ep_saude_sexo2 * 100, coef_sexo2 * 100, epa_saude_sexo2)
options(OutDec=",")
xtable( tab_saude_sexo2 , digits = 2 )
options(OutDec=".")

# Distribuição da autoavaliação da saúde | faixa etária

saude_idade2 <- svyby ( ~saude, ~idade, plano_no, svymean, deff=T)
p_saude_idade2 <- coef(saude_idade2)
ep_saude_idade2 <- c( SE(saude_idade2)[,1], SE(saude_idade2)[,2],
                    SE(saude_idade2)[,3])
coef_idade2 <- c( cv(saude_idade2)[,1], cv(saude_idade2)[,2],
                cv(saude_idade2)[,3])
epa_saude_idade2 <- c( 2.617416, 2.945925, 2.050508, 2.435220, 2.420850,
2.619457, 2.841846, 2.445762, 2.868244, 2.716352,
2.256665, 2.371590, 1.984378, 2.154551, 3.035944)

tab_saude_idade2 <- data.frame( p_saude_idade2*100, ep_saude_idade2*100 ,
                              coef_idade2 * 100, epa_saude_idade2)
options(OutDec=",")
xtable( tab_saude_idade2 , digits = 2 )
options(OutDec=".")

# Distribuição da autoavaliação da saúde | raça

saude_raca2 <- svyby ( ~saude, ~raca, plano_no, svymean, deff = T )
```

```

p_saude_raca2 <- coef(saude_raca2)
ep_saude_raca2 <- c(SE(saude_raca2)[,1], SE(saude_raca2)[,2],
                   SE(saude_raca2)[,3])
coef_raca2 <- c( cv(saude_raca2)[,1], cv(saude_raca2)[,2],
                cv(saude_raca2)[,3] )
epa_saude_raca2 <- c( 2.402818, 1.787143, 2.476649, 2.298256,
                    2.476536, 1.878600, 2.497818, 2.528648,
                    1.342747, 1.9999781, 2.3904001, 2.6406084)

tab_saude_raca2 <- data.frame( p_saude_raca2 * 100 ,
                              ep_saude_raca2 * 100, coef_raca2 * 100, epa_saude_raca2)
options(OutDec=",")
xtable( tab_saude_raca2 , digits = 2 )
options(OutDec=".")

# Distribuição da autoavaliação da saúde | nível de instrução

saude_instrucao2 <- svyby( ~saude , ~instrucao, plano_no, svymean, deff = T)
p_saude_instrucao2 <- coef(saude_instrucao2)
ep_saude_instrucao2 <- c( SE(saude_instrucao2)[,1],
                        SE(saude_instrucao2)[,2], SE(saude_instrucao2)[,3])

coef_instrucao2 <- c( cv(saude_instrucao2)[,1],
                    cv(saude_instrucao2)[,2], cv(saude_instrucao2)[,3])
epa_saude_instrucao2 <- c( 3.544001, 2.880916, 2.731846, 2.454082, 3.586035,
                        2.828873, 3.014074, 2.418755, 1.437821, 2.196413, 2.290435, 2.554165)

tab_saude_instrucao2 <- data.frame( p_saude_instrucao2 * 100,
                                   ep_saude_instrucao2 * 100, coef_instrucao2 * 100, p_saude_instrucao2)
options(OutDec=",")
xtable( tab_saude_instrucao2 , digits = 2 )
options(OutDec=".")

# Distribuição da autoavaliação da saúde | situação habitacional

saude_situ2 <- svyby ( ~ saude , ~ situ , plano_no, svymean, deff = T )
p_saude_situ2 <- coef(saude_situ2)
ep_saude_situ2 <- c(
  SE(saude_situ2)[,1], SE(saude_situ2)[,2], SE(saude_situ2)[,3])
coef_situ2 <- c(
  cv(saude_situ2)[,1], cv(saude_situ2)[,2], cv(saude_situ2)[,3])

```

```
epa_saude_situ2<-c( 2.156070,2.606816,2.294542, 2.696920, 2.316682, 2.448221)

tab_saude_situ2 <- data.frame( p_saude_situ2 * 100,
                              ep_saude_situ2*100, coef_situ2 * 100, epa_saude_situ2)
options(OutDec=",")
xtable( tab_saude_situ2 , digits = 2 )
options(OutDec=".")

##### Tabelas conjuntas #####

# Distribuição conjunta autoavaliação da saúde e sexo

saude_sex01 <- svymean( ~interaction(sexo, saude), plano_no, deff = T)
p_saude_sex01 <- coef(saude_sex01)
ep_saude_sex01 <- SE(saude_sex01)
coef_sex01 <- cv(saude_sex01)
epa_saude_sex01 <- c(1.5090, 1.1292, 1.8734, 1.5939, 1.8961, 2.3795)

tab_saude_sex01 <- data.frame( p_saude_sex01 * 100,
                              ep_saude_sex01 * 100, coef_sex01 * 100, epa_saude_sex01)
options(OutDec=",")
xtable( tab_saude_sex01 , digits = 2 )
options(OutDec=".")

# Distribuição conjunta da autoavaliação da saúde e faixa etária

saude_idade1 <- svymean( ~interaction(idade, saude), plano_no, deff=T)
p_saude_idade1 <- coef(saude_idade1)
ep_saude_idade1 <- SE(saude_idade1)
coef_idade1 <- cv(saude_idade1)
epa_saude_idade1 <-c( 1.4079, 2.1684, 1.7679, 2.1054, 1.8434, 2.6843,
                    2.1919, 2.0848, 1.9906, 2.1771, 2.5514, 2.1629, 1.8165, 1.8752, 2.7530)

tab_saude_idade1 <- data.frame( p_saude_idade1 * 100,
                              ep_saude_idade1*100, coef_idade1 * 100, epa_saude_idade1 )
options(OutDec=",")
xtable( tab_saude_idade1 , digits = 2 )
options(OutDec=".")

# Distribuição conjunta da autoavaliação da saúde e raça
saude_raca1 <- svymean( ~interaction (raca, saude), plano_no,deff =T)
```

```
p_saude_raca1 <- coef(saude_raca1)
ep_saude_raca1 <- SE(saude_raca1)
coef_raca1 <- cv(saude_raca1)
epa_saude_raca1 <- c( 3.2006, 2.7642, 2.0018, 2.7046, 2.4863,
                    1.7737, 2.6260, 3.1693, 3.1707, 3.8718,
                    0.8078, 1.8599 ,1.5425, 2.5226, 2.7333)

tab_saude_raca1 <- data.frame( p_saude_raca1 * 100,
                              ep_saude_raca1 * 100 , coef_raca1 * 100, epa_saude_raca1 )
options(OutDec=",")
xtable( tab_saude_raca1 , digits = 2 )
options(OutDec=".")

# Distribuição conjunta autoavaliação da saúde e nível de instrução
saude_instrucao1 <- svymean( ~interaction( instrucao, saude),
                           plano_no, deff = T)
p_saude_instrucao1 <- coef(saude_instrucao1)
ep_saude_instrucao1 <- SE(saude_instrucao1)
coef_instrucao1 <- cv(saude_instrucao1)
epa_saude_instrucao1 <- c( 2.9377, 2.7762, 2.7589, 2.6187, 3.6935,
                          2.8736, 3.4062, 2.4776, 1.4510, 2.2734, 2.3444, 2.2101)

tab_saude_instrucao1 <- data.frame( p_saude_instrucao1 * 100,
                                   ep_saude_instrucao1 * 100, coef_instrucao1 * 100,
                                   epa_saude_instrucao1 )
options(OutDec=",")
xtable( tab_saude_instrucao1 , digits = 2 )
options(OutDec=".")

# tabela conjunta percepção da saúde e situação habitacional
saude_situ1 <- svymean( ~interaction(situ, saude), plano_no, deff=T)
p_saude_situ1 <- coef(saude_situ1)
ep_saude_situ1 <- SE(saude_situ1)
coef_situ1 <- cv(saude_situ1)
epa_saude_situ1 <- c(2.7679, 2.4550, 2.3084, 2.7909, 2.2067, 2.4205)

tab_saude_situ1 <- data.frame( p_saude_situ1 * 100 ,
                              ep_saude_situ1 * 100 ,coef_situ1 * 100, epa_saude_situ1)
options(OutDec=",")
xtable( tab_saude_situ1 , digits = 2 )
options(OutDec=".")
```

```
##### Teste Qui-quadrado e F #####
##### correção Rao-Scott #####

### Testes de independência entre autoavaliação de saúde e Sexo

xrs_sexo <- svychisq( ~ sexo + saude, design = plano_no, statistic = "Chisq")
frs_sexo <- svychisq( ~ sexo + saude, design = plano_no, statistic = "F")

# Teste de Pearson (sem correção)

x2p.st_sexo <- xrs_sexo$statistic
x2p.gl_sexo <- xrs_sexo$parameter
data.frame(X2p_sexo = x2p.st_sexo,
           gl_sexo = x2p.gl_sexo,
           p_sexo = pchisq(x2p.st_sexo, x2p.gl_sexo, lower.tail=FALSE))

# Teste de Pearson com correção de Rao-Scott (2a. ordem)

xrs2.1_sexo <- xrs_sexo$parameter*frs_sexo$statistic
xrs2.1.df_sexo <- xrs_sexo$parameter
data.frame(X2RS_sexo = xrs2.1_sexo,
           gl_sexo=xrs2.1.df_sexo,
           p_sexo=pchisq(xrs2.1_sexo, xrs2.1.df_sexo, lower.tail=FALSE))
# Teste F com correção de Rao-Scott (2a. ordem)
frs_sexo

### Testes de independência entre autoavaliação de saúde e Faixa etária

xrs_idade <- svychisq( ~ idade + saude, design = plano_no, statistic = "Chisq")
frs_idade <- svychisq( ~ idade + saude, design = plano_no, statistic = "F")

# Teste de Pearson (sem correção)

x2p.st_idade <- xrs_idade$statistic
x2p.gl_idade <- xrs_idade$parameter
data.frame(X2p_idade = x2p.st_idade,
           gl_idade = x2p.gl_idade,
           p_idade = pchisq(x2p.st_idade, x2p.gl_idade, lower.tail=FALSE))

# Teste de Pearson com correção de Rao-Scott (2a. ordem)
```

```
xrs2.1_idade <- xrs_idade$parameter * frs_idade$statistic
xrs2.1.df_idade <- xrs_idade$parameter
data.frame(X2RS_idade = xrs2.1_idade,
           gl_idade = xrs2.1.df_idade,
           p_idade = pchisq(xrs2.1_idade, xrs2.1.df_idade, lower.tail=FALSE))

# Teste F com correção de Rao-Scott (2a. ordem)
frs_idade

### Testes de independência entre autoavaliação de saúde e Raça

xrs_raca <- svychisq( ~ raca + saude, design = plano_no, statistic = "Chisq")
frs_raca <- svychisq( ~ raca + saude, design = plano_no, statistic = "F")

# Teste de Pearson (sem correção)

x2p.st_raca <- xrs_raca$statistic
x2p.gl_raca <- xrs_raca$parameter
data.frame(X2p_raca = x2p.st_raca,
           gl_raca=x2p.gl_raca,
           p_raca=pchisq(x2p.st_raca, x2p.gl_raca, lower.tail=FALSE))

# Teste de Pearson com correção de Rao-Scott (2a. ordem)

xrs2.1_raca <- xrs_raca$parameter*frs_raca$statistic
xrs2.1.df_raca <- xrs_raca$parameter
data.frame(X2RS_raca = xrs2.1_raca,
           gl_raca=xrs2.1.df_raca,
           p_raca=pchisq(xrs2.1_raca, xrs2.1.df_raca, lower.tail=FALSE))

# Teste F com correção de Rao-Scott (2a. ordem)

frs_raca

### Testes de independência entre autoavaliação de saúde e Nível de instrucao

xrs_inst <- svychisq( ~ instrucao + saude, design = plano_no, statistic = "Chisq")
frs_inst <- svychisq( ~ instrucao + saude, design = plano_no, statistic = "F")

# Teste de Pearson (sem correção)
```

```
x2p.st_inst <- xrs_inst$statistic
x2p.gl_inst <- xrs_inst$parameter
data.frame( X2p_inst = x2p.st_inst,
            gl_inst = x2p.gl_inst,
            p_inst = pchisq(x2p.st_inst, x2p.gl_inst, lower.tail=FALSE))

# Teste de Pearson com correção de Rao-Scott (2a. ordem)

xrs2.1_inst <- xrs_inst$parameter*frs_inst$statistic
xrs2.1.df_inst <- xrs_inst$parameter
data.frame(X2RS_inst = xrs2.1_inst,
            gl_inst = xrs2.1.df_inst,
            p_inst = pchisq(xrs2.1_inst, xrs2.1.df_inst, lower.tail=FALSE))

# Teste F com correção de Rao-Scott (2a. ordem)
frs_inst

### Testes de independência entre autoavaliação de saúde e Situação habitacional

xrs_situ <- svychisq( ~ situ + saude, design = plano_no, statistic = "Chisq")
frs_situ <- svychisq( ~ situ + saude, design = plano_no, statistic = "F")

# Teste de Pearson (sem correção)

x2p.st_situ <- xrs_situ$statistic
x2p.gl_situ <- xrs_situ$parameter
data.frame( X2p_situ = x2p.st_situ,
            gl_situ = x2p.gl_situ,
            p_situ = pchisq( x2p.st_situ, x2p.gl_situ, lower.tail=FALSE))

# Teste de Pearson com correção de Rao-Scott (2a. ordem)

xrs2.1_situ <- xrs_situ$parameter * frs_situ$statistic
xrs2.1.df_situ <- xrs_situ$parameter
data.frame( X2RS_situ = xrs2.1_situ,
            gl_situ = xrs2.1.df_situ,
            p_situ = pchisq(xrs2.1_situ, xrs2.1.df_situ, lower.tail=FALSE))

# Teste F com correção de Rao-Scott (2a. ordem)
frs_situ
```

```
##### Representação gráfica das tabelas condicionais #####
```

```
# Representação gráfica da autoavaliação de saúde e sexo
```

```
t_saude_sexo <- data.frame(
  sexo = factor( rep( c( "Feminino","Masculino"),3),
                levels=c( "Masculino","Feminino")),
  saude = factor( c( rep("Boa",2), rep("Regular",2), rep("Ruim",2) ),
                 levels = c( "Ruim","Regular","Boa" ) ),
  prop = p_saude_sexo2 * 100,
  limInf = confint(saude_sexo2) [,1] * 100,
  limSup = confint(saude_sexo2) [,2] * 100)
```

```
ggplot(t_saude_sexo, aes(x=prop, y=saude)) +
  geom_point(stat="identity") + theme_bw() +
  geom_errorbarh(aes(xmin=limInf, xmax=limSup), height =0.2) +
  scale_x_continuous(limits = c(0,65), breaks = seq(0, 65, 10)) +
  labs(x="Proporção de pessoas (%)", y="Autoavaliação de saúde") +
  facet_wrap( ~ sexo, ncol=1)
```

```
# Representação gráfica da autoavaliação de saúde e faixa etária
```

```
t_saude_idade <- data.frame(
  idade = factor( rep( c( "18-29 anos", "30-59 anos", "60-64 anos",
                        "65-74 anos", "75 anos ou mais" ) , 3 ),
                 levels = c("18-29 anos", "30-59 anos", "60-64 anos",
                            "65-74 anos", "75 anos ou mais" ) ),
  saude = factor( c( rep("Boa",5), rep("Regular",5), rep("Ruim",5) ),
                 levels = c("Ruim","Regular","Boa" ) ),
  prop = p_saude_idade2 * 100,
  limInf = confint(saude_idade2) [,1] * 100,
  limSup = confint(saude_idade2) [,2] * 100)
```

```
ggplot(t_saude_idade, aes(x=prop, y=saude) ) +
  geom_point(stat="identity") + theme_bw() +
  geom_errorbarh(aes(xmin=limInf, xmax=limSup), height =0.2) +
  scale_x_continuous(limits = c(0,80), breaks = seq(0, 80, 10) ) +
  labs(x="Proporção de pessoas (%)", y="Autoavaliação de saúde") +
  facet_wrap( ~ idade, ncol=1)
```

```
# Representação gráfica da autoavaliação de saúde e raça
```

```
t_saude_raca <- data.frame( raca = factor(
```

```

rep( c( "Amarela ou Indígena", "Branca", "Parda", "Preta"),3),
levels = c( "Amarela ou Indígena", "Branca", "Parda", "Preta")),
saude = factor( c( rep("Boa",4), rep("Regular",4), rep("Ruim",4) ) ,
                levels = c("Ruim","Regular","Boa")),
prop = p_saude_raca2 * 100,
limInf = confint(saude_raca2) [, 1] * 100,
limSup = confint(saude_raca2) [, 2] * 100)

ggplot(t_saude_raca, aes(x=prop, y=saude)) +
  geom_point(stat="identity") + theme_bw() +
  geom_errorbarh(aes(xmin=limInf, xmax=limSup), height =0.2) +
  scale_x_continuous(limits = c(0,75), breaks = seq(0, 75, 10) ) +
  labs(x="Proporção de pessoas (%)", y="Autoavaliação de saúde") +
  facet_wrap( ~ raca, ncol=1)

# Representação gráfica da autoavaliação de saúde e nível de instrução
t_saude_instrucao <- data.frame( instrucao = factor(
  rep( c( "Fundamental completo ou médio incompleto",
          "Médio completo ou superior incompleto",
          "Sem instrução ou fundamental incompleto",
          "Superior completo"), 3),
  levels = c( "Sem instrução ou fundamental incompleto",
              "Fundamental completo ou médio incompleto",
              "Médio completo ou superior incompleto",
              "Superior completo" ) ),
saude = factor( c( rep("Boa",4), rep("Regular",4), rep("Ruim",4)),
                levels = c("Ruim","Regular","Boa") ),
prop = p_saude_instrucao2 * 100,
limInf = confint(saude_instrucao2) [, 1] * 100,
limSup = confint(saude_instrucao2) [, 2] * 100)

ggplot(t_saude_instrucao, aes(x=prop, y=saude)) +
  geom_point(stat="identity") + theme_bw() +
  geom_errorbarh(aes(xmin=limInf, xmax=limSup), height =0.2) +
  scale_x_continuous(limits = c(0,80), breaks = seq(0, 80, 10)) +
  labs(x="Proporção de pessoas (%)", y="Autoavaliação de saúde") +
  facet_wrap( ~ instrucao, ncol=1)

# Representação gráfica da autoavaliação de saúde e situação habitacional
t_saude_situ <- data.frame(
  situ = factor( rep( c("Rural","Urbano"), 3),

```

```
      levels = c("Urbano","Rural") ),
saude = factor( c( rep("Boa",2), rep("Regular",2), rep("Ruim",2)),
      levels = c("Ruim","Regular","Boa") ),
prop = p_saude_situ2 * 100,
limInf = confint(saude_situ2) [, 1] * 100,
limSup = confint(saude_situ2) [, 2] * 100)

ggplot(t_saude_situ, aes(x=prop, y=saude) ) +
  geom_point(stat="identity") + theme_bw() +
  geom_errorbarh(aes(xmin=limInf, xmax=limSup), height =0.3) +
  scale_x_continuous(limits = c(0,65), breaks = seq(0, 65, 10)) +
  labs(x="Proporção de pessoas (%)", y="Autoavaliação de saúde") +
  facet_wrap( ~ situ, ncol=1)
```