UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

NOEMIR DOS SANTOS SOUSA LIMA

ASSOCIAÇÃO EM TABELAS DE CONTINGÊNCIA DE DUPLA ENTRADA COM DADOS AMOSTRAIS COMPLEXOS DE COVID-19

Recife

2022

NOEMIR DOS SANTOS SOUSA LIMA

ASSOCIAÇÃO EM TABELAS DE CONTINGÊNCIA DE DUPLA ENTRADA COM DADOS AMOSTRAIS COMPLEXOS DE COVID-19

Trabalho apresentado ao Programa de Pós-graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador (a): Cristiano Ferraz

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

L732a Lima, Noemir dos Santos Sousa
Associação em tabelas de contingência de dupla entrada com dados amostrais complexos de covid-19 / Noemir dos Santos Sousa Lima. – 2022.
63 f.: il., tab.

Orientador: Cristiano Ferraz.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2022.

Inclui referências e apêndice.

1. Estatística aplicada. 2. Pesquisa sorológica. 3. Covid-19. I. Ferraz, Cristiano (orientador). II. Título.

310 CDD (23. ed.) UFPE - CCEN 2022-140

NOEMIR DOS SANTOS SOUSA

"ASSOCIAÇÃO EM TABELAS DE CONTINGÊNCIA DE DUPLA ENTRADA COM DADOS AMOSTRAIS COMPLEXOS DE COVID-19"

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 29 de junho de 2022.

BANCA EXAMINADORA

Prof. Dr. Cristiano Ferraz

DE/UFPE

Prof. Dr. Alex Dias Ramos

DE/UFPE

Prof. Dr. Hemílio Fernandes Campos Coelho

DE/UFPB

AGRADECIMENTOS

Primeiramente, a Deus, pelo dom da vida e por me permitir ultrapassar todos os obstáculos encontrados ao longo da realização deste mestrado.

Ao orientador desta dissertação, o professor Cristiano Ferraz, pela orientação prestada, por toda paciência, pelo seu incentivo, disponibilidade e apoio que sempre demonstrou.

Aos meus pais, Natália e Valdemiro, pelo apoio, pela força e pelo carinho que sempre me prestaram ao longo de toda a minha vida acadêmica.

Ao meu esposo, Evalton, por ter caminhado ao meu lado durante todo o percurso na produção desta dissertação, pela sua paciência, compreensão, pelo apoio e incentivo nas minhas escolhas.

Aos meus irmãos, Samuel, Welton, Noelma, Hélio e Hélia, que sempre vibraram cada conquista, mostrando a certeza que nunca estarei só.

Ao professor Josimar, do curso de licenciatura em Matemática da Universidade Federal do Piauí (UFPI), por ter me incentivado a participar da seleção de mestrado, pela ajuda quando fui à Recife, pela receptividade e pelo apoio.

Aos meus mestres, por todo ensinamento, sem vocês este sonho não se concretizaria.

As amigas do mestrado em Estatística da UFPE, em especial a Suelem que tanto me ajudou desde o início, agradeço por sua ajuda financeira e intelectual, também quero agradecer a Penélope e Jaciele, por serem verdadeiras companheiras de estudo.

A Fundação de Amparo à Ciência e Tecnologia do estado de Pernambuco (FACEPE), pelo apoio financeiro durante o mestrado.

Enfim, a todos que de forma direta ou indireta, me ajudaram durante este processo, sem os quais não conseguiria concluir.

RESUMO

A análise de dados epidemiológicos é essencial para o planejamento estratégico de ações de combate a surtos, epidemias e pandemias. Durante a sindemia de COVID-19 a importância de tais análises ganhou evidência devido a gravidade dos efeitos do vírus SARS-CoV-2 no Brasil e no mundo. Neste contexto, a analise estatistica de tabelas de contingencia e um dos recursos importantes para investigar relacoes entre variaveis, com destaque para estudos de associacao de fatores de risco e diagnosticos de diversos testes para a COVID-19. Em dados provenientes de amostras aleatorias simples com reposicao, ou provenientes de populacoes consideradas infinitas, a estatistica do teste qui-quadrado de Pearson, comumente usada para testar associação, converge para uma distribuicao qui-quadrado para tamanhos de amostras relativamente moderados. Todavia, em estudos nos quais os dados sao provenientes de planos amostrais complexos, a estatistica do teste de Pearson precisa de ajuste para convergir satisfatoriamente. Nao considerar essas caracteristicas de planos complexos nos testes de hipoteses pode gerar estimativas incorretas tanto dos parametros como das variancias dessas estimativas. Nesta tese, utilizamos dados de COVID-19 relativos ao estado da Paraíba, gerados pela Pesquisa Sorológica Continuar Cuidando, com o intuito de apresentar os testes de Rao-Scott e de Wald para investigar a associacao em tabelas de contingencia de dupla entrada, enfatizando a importancia de considerar corretamente o plano amostral. As analises mostraram que nao considerar os conglomerados no estudo pode levar a mudanças de decisões nos testes.

Palavras-chaves: pesquisa sorológica; teste de Rao-Scott; teste de Wald.

ABSTRACT

The analysis of epidemiological data is essential to the strategical planning of actions against outbreaks, epidemics, and pandemics. During the COVID-19 pandemic, the relevance of such types of analysis has gain evidence due to the SARS-CoV-19 virus' severity effects in Brazil and in the world. In this context, the analysis of contingency tables is one of the important resources to investigate relationships between variables, with emphasis on association between risk factors and diagnostic results for several types of COVID-19 tests. Using data generated by simple random samples with replacement, or selected from populations considered infinite, the Pearson chi-square statistic, commonly used to test for association, converges to a chi-square distribution for sample of moderate sizes. However, using data generated by complex sampling designs, the Pearson chis-square statistic needs adjustments to converge satisfactory. Do not consider the complexity of the sample design when testing hypothesis can lead to incorrect estimates of parameters as well as variances of estimates. In this dissertation, we use COVID-19 data related to the state of Paraíba, generated by the serological survey Continuar Cuidando, to introduce the Rao-Scott and Wald tests of association for two-way contingency tables, emphasizing the relevance of correctly considering the sample design. The analysis has shown that do not considering clustering in the study can lead to changes on the tests' decisions.

Keywords: serological survey; Rao-Scott test; Wald test.

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A epidemiologia tem por objetivo constatar os subgrupos da população que possuem maior risco de adoecer (GORDIS, 2017). A quantidade de dados epidemiológicos encontra-se longe da ótima. Contudo, as análises destas informações têm papel essencial no planejamento de estratégias de combate a sindemia de COVID-19 no Brasil e no mundo. O termo sindemia é utilizado para se referir a situações nas quais duas ou mais doenças de caráter epidêmico interagem de forma a causar danos maiores que a soma das doenças causariam em uma população específica (SINGER et al., 2017). Em uma sindemia o impacto das interações entre doenças é agravado por questões sociais e ambientais.

A aplicação de testes de diagnósticos de COVID-19 é fundamental para o monitoramento da gravidade da sindemia no Brasil. No entanto, por limitação de disponibilidade de recursos, apenas indivíduos sintomáticos que recorriam ao Sistema de Saúde eram testados, quando o ideal seria recorrer ao menos a testagem da população por meio de pesquisas sorológicas por amostragem probabilísticas. Em todas essas situações, a análise estatística de tabelas de contingência é estratégica para investigar relações entre variáveis, com destaque para estudos de associação de fatores de comorbidade e comportamento com diagnósticos positivos (ou negativos) de variados testes para a COVID-19. Também é possível, e, de interesse, investigar o desempenho de diversos testes comparados, por exemplo, ao padrão-ouro teste Reação em Cadeia da Polimerase (PCR). Quando o comportamento de uma das variáveis estudadas depende dos níveis de uma segunda variável, diz-se haver uma associação entre elas.

O estudo da relação entre tais variáveis, cuja natureza é categorizada, se dá na análise de tabelas de contingência, com uma variável identificada como fator de linha e outra, como fator de coluna. Tais estudos são realizados através de testes de hipóteses que investigam o quão distantes os valores observados estão, daqueles que se esperariam caso não houvesse associação.

A estatística comumente utilizada para avaliar a hipótese de não associação é a qui-quadrado de Pearson, proposta por Pearson (1900), que se baseia na comparação da diferença entre as frequências observadas e esperadas dos eventos considerados.

Em dados provenientes de amostras aleatórias simples com reposição, ou provenientes de populações consideradas infinitas, a estatística do teste qui-quadrado de Pearson converge para uma distribuição qui-quadrado para tamanhos de amostras relativamente moderados. Toda-

via, em estudos nos quais os dados são provenientes de planos amostrais complexos, planos que possuem algumas características como: estratificação, conglomerado e/ou probabilidade proporcional a uma medida de tamanho, a estatística do teste de Pearson precisa de ajuste para convergir satisfatoriamente. Não considerar essas características de planos complexos nos testes de hipóteses pode gerar estimativas incorretas tanto dos parâmetros como das variâncias dessas estimativas (PESSOA; SILVA, 1998).

Este trabalho tem por objetivo principal apresentar os principais testes de associação para tabelas de contingência de dupla entrada para dados amostrais complexos, tendo como motivação a aplicação destes testes a dados de COVID-19. O efeito de ignorar planos amostrais complexos na análise de dados de tabela de contingência, é investigado através de ilustrações de análises de dados provenientes da pesquisa sorológica Continuar Cuidando, do estado da Paraíba, cujo principal objetivo foi estimar a prevalência de casos de COVID-19 no estado. As análises foram realizadas com o auxílio do Software SAS On Demand for Academics.

Esta dissertação está dividida em sete capítulos. Após este capítulo introdutório, segue o capítulo sobre os modelos probabilísticos básicos, úteis para analisar dados de tabelas do contingência. No capítulo 3, apresentam-se e discutem-se os esquemas amostrais de amostra estratificada e amostra aleatória simples e como eles influenciam na abordagem de inferência em tabelas de contingência. No capítulo 4 são descritos os testes de associação, o baseado na diferença de duas proporções e o qui-quadrado. No capítulo 5, os testes para tabelas de contingências com dados amostrais complexos, de Rao-Scott e de Wald, são apresentados. O capítulo 6 mostra a aplicação dos testes a dados da pesquisa Sorológica Continuar Cuidando, empregando a metodologia do capítulo 5. A conclusão encontra-se no capítulo 7.

2 MODELOS PROBABILÍSTICOS

Os modelos probabilísticos têm a pretensão de descrever as principais características dos dados. Eles dependem da estrutura de como os dados foram gerados e, também, dos objetivos da análise. Nesta seção, decreveremos alguns modelos probabilísticos que serão utilizados por nós. Apresentamos suas caracterições de centralidade e dispersão.

2.1 DISTRIBUIÇÃO DE BERNOULLI

Um experimento de Bernoulli admite apenas dois resultados possíveis. Na literatura, os resultados do experimento assumem apenas dois valores: "1" ou "0". "1" significa "ocorrência do evento de interesse" e "0" "não ocorrência do evento de interesse". Definindo p a probabilidade de obter 1 e $1-p$ de obter 0. Então, se X é uma variável aleatória com esta distribuição denotamos

$$X \sim Bernoulli(p).$$

Considere X uma variável aleatória de Bernoulli com parâmetro p, $p \in (0,1)$, a função de probabilidade de X é dada por

$$\begin{cases} P(X=1) = p \\ P(X=0) = 1-p. \end{cases} \tag{2.1}$$

A média e a variância de uma variável aleatória de Bernoulli, com parâmetro p, são determinadas, respectivamente, por

$$\mathbb{E}(X) = p \quad \text{e} \quad \text{var}(X) = p(1-p).$$

2.2 DISTRIBUIÇÃO BINOMIAL

A distribuição binomial é uma das mais comuns para dados discretos. Tem-se um experimento binomial quando é realizado n ensaios independentes de Bernoulli, em que, a probabilidade de ocorrência do evento de interesse em cada ensaio é sempre igual a p, $0 < p < 1$. Portanto, se X é uma variável aleatória com essa distribuição denotamos

$$X \sim Binomial(n,p).$$

$$P(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} p^n (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \tag{2.2}$$

em que,

$$\begin{pmatrix} n \\ x \end{pmatrix} = \frac{n!}{x!(n-x)!}.$$

O parâmetro n, às vezes, é referido como tamanho da amostra. Nesse caso, a distribuição binomial é usada para modelar o número de ocorrência do evento de interesse em uma amostra n extraída com reposição de uma população de tamanho N.

A média e a variância de uma variável aleatória binomial X, com parâmetro n e p são dadas, respectivamente, por

$$\mathbb{E}(X) = np \quad \text{e} \quad \text{var}(X) = np(1-p),$$

no qual, n é o número total de ensaios e p a probabilidade de cada ensaio obter um resultado de ocorrência do evento de interesse.

2.3 DISTRIBUIÇÃO MULTINOMIAL

A distribuição multinomial é a generalização da distribuição binomial, que admite apenas dois resultados (como "ocorrência (presença) do evento (desfecho) de interesse" e "não ocorrência (ausência) do evento (desfecho) de interesse"), para mais de dois resultados, é uma distribuição discreta multivariada. Ela é usada quando é executado n repetições de um mesmo experimento com mais de dois resultados possíveis.

Sejam r resultados possíveis $(X_1, X_2, \ldots, X_r)$ mutuamente exclusivos com probabilidades associadas $(p_1, p_2, \ldots, p_r)$, de tal forma que $p_t \geq 0$, com $t = 1, 2, \ldots, r$ e $\sum_{t=1}^{r} p_t = 1$, a soma das probabilidades é igual a 1, pois os resultados possíveis são mutuamente exclusivos, logo um deve ocorrer. Desse modo, para n repetições independentes de um experimento, considere x_1 o número de vezes que o resultado X_1 ocorre, x_2 o número de vezes que o resultado X_2 ocorre, e assim por diante, com as seguintes restrições de que $0 < x_t < n$ e $\sum_{t=1}^{r} x_t = n$. O vetor $\mathbf{X} = (X_1, X_2, \ldots, X_r)$ de variáveis aleatórias segue uma distribuição multinomial com os parâmetros n e $\mathbf{p}$, onde $\mathbf{p} = (p_1, p_2, \ldots, p_r)$, denotamos por

$$\mathbf{X} \sim Multinomial(n, \mathbf{p}).$$

Possui, a seguinte, função de probabilidade conjunta

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r) = \frac{n!}{x_1!x_2!\ldots x_r!} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r}. \tag{2.3}$$

Para cada x_t específico a esperança e a variância são obtidas usando a distribuição binomial, da seguinte forma

$$\mathbb{E}(\mathbf{X}_t) = np_t \quad \text{e} \quad \text{var}(\mathbf{X}_t) = np_t(1 - p_t),$$

em que, $t = 1, 2, \ldots, r$, n é o número total de repetições do experimento e p_t a probabilidade de cada repetição obter o t-ésimo resultado bem sucedido.

2.4 DISTRIBUIÇÃO QUI-QUADRADO

Aqui, falaremos sobre uma distribuição de probabilidade de variável aleatória contínua, a distribuição qui-quadrado com v graus de liberdade. É uma das distribuições mais utilizadas na inferência estatística, muito comum em teste de hipótese, um exemplo, é o teste qui-quadrado para independência em uma tabela de contingência $L \times C$, em que L representa as linhas da tabela e C as colunas.

O quadrado de uma variável aleatoria normal padrão é uma variável aleatória qui-quadrado com 1 grau de liberdade (BUSSAB; MORETTIN, 2002). Isto é, a variável aleatória qui-quadrado pode ser vista quando v variáveis independentes com distribuição normal padrão são elevadas ao quadrado e somadas. Dessa forma,

Tomando $X = W_1^2 + W_2^2 + \cdots + W_v^2$, onde $W_1, W_2, \ldots, W_v$ são variáveis aleatórias normais padrão mutuamente independentes. Logo, se X é uma variável aleatória com esta distribuição denotamos

$$X \sim \chi_v^2.$$

Considere X uma variável aletória qui-quadrado com v graus de liberdade, $v \in \mathbb{N}$. A função densidade de probabilidade de X é representada por

$$f(x) = \begin{cases} \frac{x^{\frac{v}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{v}{2}) 2^{\frac{v}{2}}}, & x > 0, \\ 0, & x < 0. \end{cases} \tag{2.4}$$

no qual, $\Gamma(.)$ é a função gama.

A média e a variância da variável aleatória X com distribuição qui-quadrado com v graus de liberdade, são dadas, respectivamente, por

$$\mathbb{E}(X) = v \quad \text{e} \quad \text{var}(X) = 2v.$$

3 TABELAS DE CONTINGÊNCIA EM POPULAÇÕES FINITAS

Tabelas de contingência bidimensional são uma forma de apresentar frequências de ocorrências de resultados simultâneos relativos a duas variáveis categorizadas. Uma tabela de contingência possui uma linha para cada nível de uma primeira variável e uma coluna para cada nível de uma segunda. Define-se como uma "casela" da tabela, o resultado observado para a linha L e coluna C, isto é, a frequência de ocorrência do nível L, e do nível C da variável coluna, simultaneamente.

A tabela de contingência $L \times C$, com L linhas e C colunas, é o caso geral de uma tabela de contingência, em que, a variável de linha possui L níveis e a variável de coluna C níveis. Para efeito deste capítulo vamos nos atentar a uma tabela 2×2, como a Tabela 1 abaixo, na qual Y_{lc} representa o número de elementos de uma população finita, de tamanho $N = Y_{++}$, classificados no nível l do Fator A, $l \in \{1, 2\}$ e nível c do Fator B, $c \in \{1, 2\}$, simultaneamente.

Tabela 1 – Frequências absolutas

Frequências absolutas			
	Fator B		
Fator A	Nível 1	Nível 2	Total
Nível 1	Y_{11}	Y_{12}	Y_{1+}
Nível 2	Y_{21}	Y_{22}	Y_{2+}
Total	Y_{+1}	Y_{+2}	$Y_{++} = N$

Fonte: Elaborada pela autora (2022)

As marginais Y_{1+} e Y_{2+} (marginais de linha) representam o número de elementos da população no nível 1 do Fator A, e nível 2 do Fator A, respectivamente. E as marginais Y_{+1} e Y_{+2} (marginais de coluna) representam o número de elementos da população no nível 1 do Fator B, e nível 2 do Fator B, respectivamente, de tal forma que:

$$Y_{1+} = Y_{11} + Y_{12}; \quad Y_{2+} = Y_{21} + Y_{22}; \quad \text{e}$$

$$Y_{+1} = Y_{11} + Y_{21}; \quad Y_{+2} = Y_{12} + Y_{22}.$$

Os dados de uma tabela de contingência podem ser obtidos de diversas formas, e a compreensão de como os dados são obtidos é importante para saber o tipo de análise adequada a ser

empregada. Considerando a sua natureza, estudos podem ser classificados em dois tipos: estudos observacionais e estudos experimentais. Em estudos observacionais o investigador apenas observa o efeito da exposição dos sujeitos a determinadas condições na natureza. Nos estudos experimentais, no entanto, o investigador interfere através do uso de aleatorização para determinar quais sujeitos serão expostos a determinadas condições (STOKES; DAVIS; KOCH, 2012).

O tipo de estudo, bem como o esquema de amostragem que gera os dados determinam as condições de observação do fenômeno, e como consequência, a forma de análise estatística que deve ser empregada para realizar inferência.

Em tabelas de contingência de dupla entrada frequentemente o interesse recai em investigar a relação entre os fatores de classificação (variáveis) através de testes de hipóteses. Assim sendo, são apresentados os principais aspectos teóricos relacionados a testes de associação em tabelas de contingência de dupla entrada para dados observacionais obtidos por pesquisas amostrais. A Tabela 2 representa frequências relativas conjuntas das informações da Tabela 1, de tal forma que $p_{lc} = Y_{lc}/Y_{++}$.

Tabela 2 – Frequências relativas em relação ao total

Frequências relativas			
	Fator B		
Fator A	Nível 1	Nível 2	Total
Nível 1	p_{11}	p_{12}	p_{1+}
Nível 2	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

Fonte: Elaborada pela autora (2022)

À medida que a população N cresce e os eventos de classificação de acordo com os níveis dos Fatores A e B são equiprováveis, as frequências relativas p_{lc} descritas na Tabela 2 podem ser interpretadas como probabilidades conjuntas. As marginais p_{l+} e p_{+c} representam probabilidades de observar um elemento da população no nível l do Fator A, e nível c do Fator B, respectivamente, de tal forma que:

$$p_{l+} = p_{l1} + p_{l2}; \quad p_{+c} = p_{1c} + p_{2c}; \quad \text{e}$$

$$p_{1+} + p_{2+} = p_{+1} + p_{+2} = 1.$$

Investigar possíveis relações entre os Fatores A e B nesse contexto, equivale a verificar se existe algum padrão de comportamento das probabilidades conjuntas que variem com os níveis desses fatores. Duas perspectivas de análises são possíveis: quando os Fatores A e B são variáveis-resposta, e quando um dos Fatores é variável explicativa e o outro variável-resposta:

(i) Relação entre Fatores A e B que são variáveis-resposta:

Quando a classificação de um elemento da população nos níveis do Fator A independe da classificação desse mesmo elemento em relação aos níveis do Fator B, espera-se observar na Tabela 2 a relação descrita pela hipótese de independência dos Fatores:

$$H_0^{independência} : p_{lc} = p_{l+}p_{+c}. \tag{3.1}$$

(ii) Relação entre um Fator A, que é variável explicativa, e um Fator B que é variável-resposta:

Quando a classificação de um elemento do Fator A de acordo com os níveis do Fator B, é realizada na mesma frequência relativa para todos os níveis do fator A, espera-se observar na Tabela 2 a relação descrita pela hipótese de homogeneidade:

$$H_0^{homogeneidade} : \frac{p_{1c}}{p_{1+}} = \frac{p_{2c}}{p_{2+}} = p_{+c}. \tag{3.2}$$

A hipótese de homogeneidade é realizada sobre probabilidades condicionais $p_{c|l}$ descritas na Tabela 3.

Tabela 3 – Frequências relativas em relação as linhas

Frequências relativas			
	Fator B		
Fator A	Nível 1	Nível 2	Total
Nível 1	$p_{1\|1} = p_{11}/p_{1+}$	$p_{2\|1} = p_{12}/p_{1+}$	p_{1+}
Nível 2	$p_{1\|2} = p_{21}/p_{2+}$	$p_{2\|2} = p_{22}/p_{2+}$	p_{2+}
Total	p_{+1}	p_{+2}	1

Fonte: Elaborada pela autora (2022)

Testes de hipóteses em tabelas de contingência examinam até que ponto existem evidências nos dados observados contra as hipóteses nulas de independência e homogeneidade descritas acima.

3.1 PLANOS AMOSTRAIS E MODELOS PROBABILÍSTICOS

Considerando a perspectiva de usar uma pesquisa amostral para investigar as hipóteses de independência e homogeneidade, a Tabela 1 corresponderia a dados censitários de uma população de tamanho N. Uma amostra $S \subset U$, de tamanho $n \leq N$, é selecionada de U segundo um plano amostral probabilístico determinado, plano em que os elementos da amostra são classificados aletoriamente e cada elemento tem a mesma probabilidade de estar na amostra, através do acesso a todos os elementos da população, seja por lista ou cadastro, por exemplo. Nesta seção descrevemos como o plano amostral que gera S influencia na estratégia de testar hipóteses e consequentemente, no que é possível estudar sobre a relação entre os Fatores A e B.

Em diversos livros de análise de dados categorizados, como Paulino e Singer (2006), descreve-se a linha de análise a partir dos modelos probabilísticos considerados por nós nas subseções 2.1 até 2.4. Neste texto, evidência é dada ao plano amostral empregado, que determina quais modelos probabilísticos seriam razoáveis empregar para realizar a análise. Os planos amostrais considerados nessa apresentação são com restrição (amostra aleatória estratificada) ou de aleatorização irrestrita (amostra aleatória simples). Descreve-se primeiro como um plano de amostragem estratificada influencia na análise de tabelas de contingência pelo fato deste induzir as condições necessárias para realização de inferência a partir da distribuição Binomial. Em seguida, apresenta-se o esquema amostral de amostra aleatória simples, que induz as condições de análise a partir de um modelo Multinomial.

3.1.1 Amostra estratificada

Quando informações a respeito dos níveis do Fator A são conhecidas e disponíveis no cadastro usado para planejar a amostra, é possível selecionar a amostra $S \subset U$ utilizando um plano de amostragem estratificada no qual cada nível do Fator A corresponde a um estrato, e amostras aleatórias simples são selecionadas de forma independentes, de cada um deles. Neste caso, o Fator A serve de variável explicativa, e o Fator B, de variável resposta.

Considere uma amostra aleatória estratificada, de tamanho n, de tal forma que uma amostra aleatória simples de tamanho n_1 seja selecionada dentre os elementos do nível 1 do Fator A, e uma outra amostra aleatória simples de tamanho n_2 seja selecionada, de forma independente da primeira, dentre os elementos do nível 2 do mesmo Fator A, com $n_1 + n_2 = n$. Dessa forma,

os dados amostrais gerados podem ser resumidos pela Tabela 4, na qual as marginais-linha, $y_{1+} = n_1$ e $y_{2+} = n_2$, são fixadas de antemão, pelo esquema amostral.

Tabela 4 – Frequências amostrais sob amostra estratificada pelos níveis do Fator A

Frequências amostrais			
Fator A (estrato)	Fator B (Resposta) Nível 1	Nível 2	Total
Nível 1	y_{11}	y_{12}	$y_{1+} = n_1$
Nível 2	y_{21}	y_{22}	$y_{2+} = n_2$
Total	y_{+1}	y_{+2}	$y_{++} = n$

Fonte: Elaborada pela autora (2022)

Para efeito de desenvolvimento de um contexto, imagine que a população pesquisada seja de habitantes de uma determinada região geográfico-administrativa, como uma Unidade da Federação. Imagine ainda que o Fator A seja município, com o nível 1 representando a capital do estado, e o nível 2, os demais municípios.

Denote por $U_1 = \{1, 2, \ldots, N_1\}$ o conjunto (estrato 1) de pessoas residentes na capital, e por $U_2 = \{1, 2, \ldots, N_2\}$ o conjunto (estrato 2) de pessoas residentes nos demais municípios do estado considerado.

O conjunto de todos as pessoas residentes no estado considerado, cujo tamanho é $N = N_1 + N_2$, pode ser denotado por $U = U_1 \cup U_2$. Uma amostra $S_1 \subset U_1$, de tamanho n_1, é selecionada de U_1 usando amostra aleatória simples. Uma amostra $S_2 \subset U_2$, de tamanho n_2, é selecionada de U_2 usando amostra aleatória simples. S_1 e S_2 são selecionadas de forma independente, e $S = S_1 \cup S_2$.

Cada uma das n_1 pessoas da amostra selecionada de U_1, bem como cada uma das n_2 pessoas da amostra selecionada de U_2, ao serem investigados, são classificados de acordo com o Fator B, que possui dois níveis.

Em adição suponha, para efeito de ilustração, que o Fator B seja um fator de comportamento relacionado ao enfrentamento da sindemia de COVID-19, com respostas possíveis: usa (nível 1) ou não usa (nível 2) máscara sempre que sai de casa. Isso implicaria que as informações da Tabela 4 poderiam ser melhor descritas pela Tabela 5 abaixo:

Tabela 5 – Comportamento preventivo por estrato

Comportamento preventivo por estrato			
	Usa máscara sempre que sai de casa?		
Estrato	Sim	Não	Total
Capital	y_{11}	y_{12}	$y_{1+} = n_1$
Demais	y_{21}	y_{22}	$y_{2+} = n_2$
Total	y_{+1}	y_{+2}	$y_{++} = n$

Fonte: Elaborada pela autora (2022)

Os dados da Tabela 5 são usados para estimar as probabilidades condicionais da Tabela 3, a partir das frequências relativas amostrais por linha, dadas pela Tabela 6.

Tabela 6 – Comportamento preventivo por estrato

Comportamento preventivo por estrato			
	Usa máscara sempre que sai de casa?		
Estrato	Sim	Não	Total
Capital	$p_{1\|1} = y_{11}/y_{1+}$	$p_{2\|1} = y_{12}/y_{1+}$	1
Demais	$p_{1\|2} = y_{21}/y_{2+}$	$p_{2\|2} = y_{22}/y_{2+}$	1
Total	p_{+1}	$(1 - p_{+1})$	

Fonte: Elaborada pela autora (2022)

Neste contexto de estudo, a hipótese de homogeneidade (6) pode ser escrita como:

$$H_0 : p_{1|1} = p_{1|2} \Leftrightarrow p_{1|1} - p_{1|2} = 0 \tag{3.3}$$

Para simplificar notação, considere $p_1 = p_{1|1}$, e $p_2 = p_{1|2}$ como a probabilidade de uma pessoa usar máscara sempre que sai na capital e nos demais municípios, respectivamente. Testar a hipótese de homogeneidade significa investigar se há evidência suficiente para refutar a ideia de que a probabilidade de uma pessoa responder que usa sempre máscara ao sair de casa é a mesma para quem mora na capital e nos demais municípios.

A derivação de testes de hipóteses nesse contexto fundamenta-se na distribuição Binomial, e depende das seguintes condições:

- C1: Em cada estrato, a amostra aleatória simples é selecionada com reposição, ou sem reposição com fração amostral negligível;
- C2: A unidade amostral (de análise) é uma pessoa;
- C3: A probabilidade de usar sempre a máscara seja a mesma dentre todos os que moram na capital;
- C4: A probabilidade de usar sempre a máscara seja a mesma dentre todos os que moram nos demais municípios;
- C5: A decisão de usar máscara sempre que sai de casa é independente de pessoa para pessoa.

Defina: $Y_1 =$ número de pessoas da capital que respondem que usam máscara sempre que saem de casa;

$Y_2 =$ número de pessoas dos demais municípios que respondem que usam máscara sempre que saem de casa.

Sob as condições C1 a C5,

$$Y_1 \sim Binomial(n_1, p_1);$$

$$Y_2 \sim Binomial(n_2, p_2).$$

3.1.2 Amostra aleatória simples

Quando uma amostra $S \subset U$ é selecionada utilizando um plano de amostragem aleatória simples, e cada elemento da amostra é classificado simultaneamente de acordo com os níveis do Fator A e do fator B, ambos os fatores servem como variáveis-resposta.

Considere uma amostra aleatória simples, de tamanho n, selecionada da população U, de tamanho N. Dessa forma, os dados amostrais gerados podem ser resumidos pela Tabela 7, na qual apenas o total geral, $y_{++} = n$, é fixado de antemão pelo esquema amostral.

Tabela 7 – Frequências amostrais sob amostra aleatória simples

Frequências amostrais			
Fator A (Resposta)	Fator B (Resposta) Nível 1	Nível 2	Total
Nível 1	y_{11}	y_{12}	$y_{1+} = n_1$
Nível 2	y_{21}	y_{22}	$y_{2+} = n_2$
Total	y_{+1}	y_{+2}	$y_{++} = n$

Fonte: Elaborada pela autora (2022)

Para efeito de desenvolvimento de um contexto, imagine que a população pesquisada seja de habitantes de uma determinada região geográfico-administrativa, como uma Unidade da Federação. Imagine ainda que o Fator A seja a informação de que a pessoa tem que trabalhar ou não durante o período de isolamento social, e o Fator B seja a resposta sobre se usa (nível 1) ou não usa (nível 2) máscara sempre que sai de casa. Logo, a Tabela 7 poder ser mais bem contextualizada pela Tabela 8 abaixo.

Tabela 8 – Comportamento durante o isolamento

Comportamento preventivo por estrato			
Trabalha durante o isolamento?	Usa máscara sempre que sai de casa? Sim	Não	Total
Sim	y_{11}	y_{12}	$y_{1+} = n_1$
Não	y_{21}	y_{22}	$y_{2+} = n_2$
Total	y_{+1}	y_{+2}	$y_{++} = n$

Fonte: Elaborada pela autora (2022)

Os dados da Tabela 8 são usados para estimar as probabilidades conjuntas da Tabela 2, a partir das frequências relativas amostrais dadas pela Tabela 9.

Tabela 9 – Frequências relativas de comportamento durante o isolamento

Frequências relativas			
Trabalha durante o isolamento?	Usa máscara sempre que sai de casa? Sim	Não	Total
Sim	p_{11}	p_{12}	p_{1+}
Não	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

Fonte: Elaborada pela autora (2022)

Neste contexto de estudo, a hipótese de independência (5) pode ser escrita como:

$$H_0 : p_{lc} = p_{l+}p_{+c}. \tag{3.4}$$

Testar a hipótese de independência significa investigar se há evidência suficiente para refutar a ideia de que a probabilidade de uma pessoa responder que usa sempre máscara ao sair de casa não é influenciada pelo fato da mesma ter que trabalhar (ou não) durante medidas de isolamento social.

A derivação de testes de hipóteses nesse contexto fundamenta-se na distribuição Multinomial, e depende das seguintes condições:

- C1: A amostra aleatória simples é selecionada com reposição, ou sem reposição com fração amostral negligível;
- C2: A unidade amostral (de análise) é uma pessoa;
- C3: A probabilidade de ter que trabalhar durante o isolamento seja a mesma dentre todas as pessoas do estado;
- C4: A probabilidade de usar sempre a máscara seja a mesma dentre todas as pessoas do estado;
- C5: O fato de ter que trabalhar durante o isolamento deve ser independente de pessoa para pessoa;
- C6: A decisão de usar máscara sempre que sai de casa é independente de pessoa para pessoa.

Defina quatro categorias de acordo com os resultados possíveis de classificação nos Fatores A e B:

Categoria (i)	Casela (l, c)
1	$(1, 1)$
2	$(1, 2)$
3	$(2, 1)$
4	$(2, 2)$

Fonte: Elaborada pela autora (2022)

Defina ainda:

Y_i= número de pessoas classificadas na categoria i dentre n pessoas na amostra;

p_i= probabilidade de uma pessoa ser classificada na categoria i;

$\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)^T$; e $\mathbf{p} = (p_1, p_2, p_3, p_4)^T$.

Sob as condições C1 a C6,

$$\mathbf{Y} \sim Multinomial(n, \mathbf{p}).$$

4 TESTE DE ASSOCIAÇÃO

Neste capítulo apresentamos primeiro o teste para diferença entre duas proporções e em seguida, o teste qui-quadrado de Pearson como uma extensão do primeiro. Os exemplos estão relacionados ainda a tabela 2×2 para simplificar a compreensão dos conceitos.

4.1 TESTE PARA A DIFERENÇA DE DUAS PROPORÇÕES

Considere Y_{11}, total de elementos classificados na casela $(1,1)$, e Y_{21}, total de elementos classificados na casela $(2,1)$, duas variáveis aleatóriais com distribuição Binomial com parâmetros n_1, p_1 e n_2, p_2, sendo p_1 a proporção de um elemento ser classificado no nível 1 do Fator A de acordo com o nível 1 do Fator B, p_2 a proporção dele ser classificado no nível 2 do Fator A de acordo com o nível 1 o Fator B, n_1 a marginal de linha 1 e n_2 a de linha 2. A diferença das duas proporções $p_1 - p_2$ está entre -1 e $+1$ e é igual a zero quando $p_1 = p_2$, quer dizer, quando o elemento é classificado independente do nível do Fator A.

A hipótese nula de que as proporções p_1 e p_2 são iguais é dada por

$$H_0 : p_1 - p_2 = 0,$$

e a hipótese alternativa é representada por

$$H_0 : p_1 - p_2 \neq 0.$$

Para amostras grandes usa-se o teste Z baseado na estatística Z, que segue a distribuição normal padrão sob a hipótese nula. O teste é construído com base nas propriedades estatísticas do estimador

$$\hat{p}_1 - \hat{p}_2 = p_1 - p_2,$$

no qual, $p_1 = \frac{Y_{11}}{n_1}$ e $p_2 = \frac{Y_{21}}{n_2}$.

- $\mathbb{E}(\hat{p}_1 - \hat{p}_2) = p_1 - p_2,$
- $\mathrm{var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$

Quando a hipótese nula é verdadeira, a variância acima pode ser estimada utilizando os dados da linha 1 e 2, como segue:

$$\widehat{\text{var}}(\hat{p_1} - \hat{p_2}|H_0) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}, \quad \text{com} \quad p = \frac{Y_{+1}}{n_1 + n_2} = \frac{Y_{11} + Y_{21}}{n_1 + n_2}.$$

A estatística Z é representada por:

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}},$$

Quando os valores de n_1 e n_2 são suficientemente grandes, satisfazendo as condições $min\{n_1, n_2\} \times p \geq 5$ e $min\{n_1, n_2\} \times (1-p) \geq 5$, tem-se que

$$Z \sim Normal(0, 1),$$

e o valor-p é determinado, de forma aproximada, pela probabilidade

$$P(|Z| > z) = P(Z < -z) + P(Z < +z).$$

Valores-p muito pequenos, usualmente menores que 0,05, fornecem evidência suficiente para rejeitar a hipótese nula.

4.2 TESTES QUI-QUADRADO

O teste qui-quadrado de Pearson é utilizado para analisar hipóteses de independência e de homogeneidade. A estatística de teste qui-quadrado mede a distância entre valores observados e esperados sob a hipótese nula de independência, e de homogeneidade (LOHR, 2010). Ela se origina da estatística Z, dado que, Z^2 segue uma distribuição qui-quadrado, e representa uma extensão para comparação de mais de duas proporções. Além disso, possui distribuição de referência qui-quadrado com $(L-1)(C-1)$ graus de liberdade (PESSOA; SILVA, 1998), L total de linhas e C total de colunas na tabela.

4.2.1 Teste qui-quadrado de independência

Para o teste independência, considere os dados da Tabela 9 que possui as frequências relativas de comportamento durante o isolamento, onde o Fator A (Trabalha durante o isolamento?) e Fator B (Usa máscara sempre que sai de casa?) são considerados variáveis-resposta. O teste qui-quadrado de Pearson é uma opção para investigar se há evidência para negar a

ideia de que a probabilidade de uma pessoa responder que usa máscara ao sair de casa não é influenciada pelo fato da mesma ter que trabalhar (ou não) durante as medidas de isolamento social.

A hipótese de independência (3.4), nesse contexto, é representada por

$$H_0 : p_{lc} = p_{l+}p_{+c} \quad l = 1, 2 \quad e \quad c = 1, 2.$$

A hipótese nula afirma que as variáveis trabalha durante o isolamento e usa máscara sempre que sai de casa, na Tabela 9, não estão relacionadas. Assim, é possível escrever H_0 : As variáveis trabalha durante o isolamento e usa máscara sempre que sai de casa são independentes.

A hipótese nula é definida pelo motivo de que quando as variáveis são independentes a probabilidade conjunta é igual a multiplicação das probabilidades das marginais.

Sob a hipótese de independência das variáveis em estudo as frequências esperadas em cada casela são calculadas da seguinte forma $\mathbb{E}_{lc} = np_{lc}$, se H_0 é verdadeiro, então $\mathbb{E}_{lc} = np_{l+}p_{+c}$, logo $\mathbb{E}_{lc}$ pode ser estimada por

$$\hat{\mathbb{E}}_{lc} = n\hat{p}_{l+}\hat{p}_{+c} = n\frac{y_{l+}}{n}\frac{y_{+c}}{n} = \frac{y_{l+}y_{+c}}{n},$$

com $\hat{p}_{lc} = y_{lc}/n$ a proporção estimada da casela (l, c), $\hat{p}_{+c} = \sum_{l=1}^{L} \hat{p}_{lc}$ proporção estimada na coluna c e $\hat{p}_{l+} = \sum_{c=1}^{C} \hat{p}_{lc}$ proporção estimada na linha l. A Tabela 10 tem os valores esperados para as observações da Tabela 8.

Tabela 10 – Frequências esperadas de comportamento durante o isolamento

Frequências esperadas			
Trabalha durante o isolamento?	Usa máscara sempre que sai de casa? Sim	Não	Total
Sim	$\mathbb{E}_{11} = y_{1+}y_{+1}/n$	$\mathbb{E}_{12} = y_{1+}y_{+2}/n$	y_{1+}
Não	$\mathbb{E}_{21} = y_{2+}y_{+1}/n$	$\mathbb{E}_{22} = y_{2+}y_{+2}/n$	y_{2+}
Total	y_{+1}	y_{+2}	n

Fonte: Elaborada pela autora (2022)

A estatística do teste qui-quadrado de Pearson para a hipótese de independência é dada por

$$X^2(I) = \sum_{l=1}^{L}\sum_{c=1}^{C} \frac{(y_{lc} - \hat{\mathbb{E}}_{lc})^2}{\hat{\mathbb{E}}_{lc}} = n\sum_{l=1}^{L}\sum_{c=1}^{C} \frac{(\hat{p}_{lc} - \hat{p}_{l+}\hat{p}_{+c})^2}{\hat{p}_{l+}\hat{p}_{+c}} \tag{4.1}$$

Conforme a hipótese nula de independência, espera-se que as frequências observadas (y_{lc}) em cada casela não sejam muito diferentes das frequências esperadas($\mathbb{E}_{lc}$). Sob a hipótese nula, a estatística $X^2(I)$ tem distribuição de referência qui-quadrado com $(L-1)(C-1) = 1$ grau de liberdade (PESSOA; SILVA, 1998), ou seja, $(2-1)(2-1) = 1$ grau de liberdade.

4.2.2 Teste qui-quadrado de homogeneidade

Considerando o teste homogeneidade, admita os dados da Tabela 6 que possui as frequências relativas amostrais por linha do comportamento preventivo por estrato, onde o Fator A (Estrato), é considerado variável explicativa, e Fator B (Usa máscara sempre que sai de casa?) variável-resposta. O teste qui-quadrado de Pearson é uma escolha para investigar se há evidência suficiente para negar a ideia de que a probabilidade de uma pessoa responder que usa máscara ao sair de casa não é a mesma para quem mora na capital e nos demais municípios.

A hipótese nula de homogeidade (7) expressa por

$$H_0 : p_{1|1} = p_{1|2}$$

onde, $p_{1|1} = y_{11}/y_{1+}$ é a probabilidade de uma pessoa usar máscara sempre que sai na capital e $p_{1|2} = y_{21}/y_{2+}$ a probabilidade de uma pessoa usar máscara sempre que sai nos demais municípios.

A hipótese nula diz que a probabilidade de uma pessoa da capital responder que usa máscara sempre que sai de casa é a mesma de uma pessoa que mora nos demais municípios. Logo, é possível escrever H_0 : Os estratos capital e demais municípios são homogêneos.

A estatística de teste de Pearson para a hipótese de homogeneidade é determinada por

$$X^2(H) = \sum_{l=1}^{2}\sum_{c=1}^{2} \frac{(y_{lc} - \hat{\mathbb{E}}_{lc})^2}{\hat{\mathbb{E}}_{lc}} = \sum_{l=1}^{2}\sum_{c=1}^{2} \frac{y_{l+}(\hat{p}_{lc} - \hat{p}_{+c})^2}{\hat{p}_{+c}}, \tag{4.2}$$

no qual, $y_{l+} = \sum_{c=1}^{2} y_{lc}$ marginal de linha l, $l = 1, 2$, $\hat{p}_{lc} = y_{lc}/y_{l+}$ proporção estimada na casela (l, c), $\hat{p}_{l+} = \sum_{c=1}^{2} \hat{p}_{lc}$ proporção estimada na linha l, $l = 1, 2$ e $\hat{p}_{+c} = \sum_{l=1}^{2} \hat{p}_{lc}$ proporção estimada na coluna c, $c = 1, 2$.

Sob a hipótese nula de homogeneidade a estatística $X^2(H)$, também, tem distribuição de referência qui-quadrado com $(L-1)(C-1)$ graus de liberdade (PESSOA; SILVA, 1998), isto é, $(2-1)(2-1) = 1$ grau de liberdade.

5 TABELAS DE CONTINGÊNCIA DE DADOS AMOSTRAIS COMPLEXOS

Os dados amostrais de uma tabela de contingência são complexos quando são coletados a partir de planos amostrais que envolvem estratificação, conglomeração e/ou probabilidades desiguais de seleção. O plano amostral, ou desenho amostral, pode interferir tanto nas probabilidades estimadas das caselas quanto nos testes qui-quadrado de independência e homogeneidade, uma vez que, o desenho possui agrupamentos e não há mais a amostragem aleatória, que resultaria em uma distribuição qui-quadrado (LOHR, 2010). Não considerar o plano amostral adotado na pesquisa pode induzir, aliás, a resultados errôneos de não-rejeição da hipótese nula.

Na tabela 11, Y_{lc} é o número de elementos classificados no nível (linha) L da variável 1 e nível (coluna) C da variável 2 sobre as quais cada elemento da população tem sua relação de pertencimento. A estrutura populacional, com fatores cruzados, não necessariamente coincide com a estrutura amostral observada. Em amostras geradas por planos amostrais complexos, outros fatores, que não variável 1 e 2, podem corresponder a estratos e/ou conglomerados, e devem ser considerados para efeito de análise.

Tabela 11 – Estrutura de uma tabela de contingência com dupla entrada

Tabela de contingência com dupla entrada							
	Variável 2						
Variável 1	1	2	...	c	...	C	Total
1	Y_{11}	Y_{12}	...	Y_{1c}	...	Y_{1C}	Y_{1+}
2	Y_{21}	Y_{22}	...	Y_{2c}	...	Y_{2C}	Y_{2+}
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
l	Y_{l1}	Y_{l2}	...	Y_{lc}	...	Y_{lC}	Y_{l+}
$\vdots$	$\vdots$	$\vdots$	.	$\vdots$	.	$\vdots$	$\vdots$
L	Y_{L1}	p_{L2}	...	Y_{Lc}	...	Y_{LC}	Y_{L+}
Total	Y_{+1}	Y_{+2}	...	Y_{+c}	...	Y_{+C}	$Y_{++} = N$

Fonte: Elaborada pela autora (2022)

Considere uma população estruturada em H estratos, cada qual contendo um total de N_h conglomerados, de tamanho M_{hi}. Defina U_h como o conjunto de todos os conglomerados do estrato h e U_{hi} como o conjunto de todos os indivíduos pertencentes ao conglomerado i no estrato h. É possível escrever então as quantidades populacionais da tabela 11 como:

$$Y_{lc} = \sum_{h=1}^{H} \sum_{i \in U_h} \sum_{j \in U_{hi}} \delta_{hij}(l,c) \tag{5.1}$$

com $\delta_{hij}(l,c) = 1$, se o indivíduo j do conglomerado i no estrato h pertencer a categoria (l,c) da tabela 1; e zero, caso contrário. A notação ressalta o fato dos parâmetros serem totais populacionais de domínios de interesse. Os totais das caselas (l,c) são estimados de forma centrada utilizando Horvitz-Thompson (HORVITZ; THOMPSON, 1952):

$$\hat{Y}_{lc} = \sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{\delta_{hij}(l,c)}{\pi_{hij}} = \sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \delta_{hij}(l,c) w_{hij} \tag{5.2}$$

onde π_{hij} e $w_{hij} = 1/\pi_{hij}$ correspondem a probabilidade de inclusão de primeira ordem e peso amostral do indivíduo j do conglomerado i no estrato h, respectivamente. Ainda em (5.2), S_h é o conjunto de conglomerados selecionados no estrato h, e S_{hi} é o conjunto de indivíduos selecionados para a amostra do conglomerado i no estrato h. Da mesma forma, as estimativas de totais de linha, de coluna e totais gerais são representadas como

$$\hat{Y}_{l+} = \sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \delta_{hij}(l,+) w_{hij} \quad \hat{Y}_{+c} = \sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} \delta_{hij}(+,c) w_{hij} \quad \hat{N} = \sum_{h=1}^{H} \sum_{i \in S_h} \sum_{j \in S_{hi}} w_{hij} \tag{5.3}$$

A variância de $\hat{Y}_{lc}$ pode ser estimada utilizando:

$$\widehat{\text{var}}(\hat{Y}_{lc}) = \sum_{h=1}^{H} \widehat{\text{var}}_h(\hat{Y}_{lc}) \tag{5.4}$$

com

$$\widehat{\text{var}}_h(\hat{Y}_{lc}) = \sum_{h=1}^{H} \left\{ \frac{S_h(1-f_h)}{S_h - 1} \left[\sum_{i \in S_h} \sum_{j \in S_{hi}} \delta_{hij}(l,c) w_{hij} - \bar{\delta}_{hi}(l,c) \right]^2 \right\}$$

e

$$\bar{\delta}_{hi}(l,c) = \sum_{i \in S_h} \sum_{j \in S_{hi}} \frac{\delta_{hij}(l,c) W_{hij}}{S_h}.$$

A matriz de covariância das estimativas de frequência da casela da tabela é uma matriz $lc \times lc$ que contém as covariâncias de frequência da casela par a par. Denote $\mathbf{V}(\hat{\mathbf{N}})$ a matriz de covariância e $\widehat{\text{cov}}(\hat{Y}_{lc}, \hat{Y}_{ab})$ as covariâncias par a par, para $l = 1, \ldots, L$, $c = 1, \ldots, C$, $a = 1, \ldots, L$ e $b = 1, \ldots, C$. A covariância entre as estimativas de frequência para as caselas da tabela (l,c) e (a,b) é estimada conforme

$$\widehat{\text{cov}}(\hat{Y}_{lc}, \hat{Y}_{ab}) = \sum_{h=1}^{H} \left(\frac{S_h(1-f_h)}{S_h - 1} \sum_{i=1}^{S_h} (S_{lc}^{hi} - \bar{S}_{lc}^{h})(S_{ab}^{hi} - \bar{S}_{ab}^{h}) \right).$$

Em planos amostrais complexos os dados possuem agrupamentos, tais como, conglomerados e/ou estratos. As estatísticas de teste para analisar a independência e homogeneidade da variável 1 e 2 precisam ser ajustadas, pois, os procedimentos de teste são complicados por causa dos efeitos de agrupamento (PESSOA; SILVA, 1998). Dessa forma, conglomerados e/ou estratos, necessitam ser considerados na estatística de teste.

Assim, uma das maneiras de obter inferências válidas utilizando amostras complexas é efetuando correções na estatística de teste de Pearson, como os ajustes de Rao-Scott. Ou, então, usando outras estatísticas de teste que já incorporem o plano amostral, tais como a estatística de Wald. Realizando essas correções a estatística terá a mesma distribuição de referência que a obtida para o caso de amostragem aleatória simples, a qui-quadrado com $(L-1)(C-1)$ graus de liberdade (PESSOA; SILVA, 1998), e, consequentemente, obter inferências válidas.

Para o teste de independência considere uma população N na tabela 11, onde a proporção de cada casela é calculada em relação ao total da população N, então $p_{lc} = Y_{lc}/N$ é a proporção da população na linha l e coluna c, as porporções das marginais são representadas por $p_{l+} = Y_{l+}/N$ proporção da população na linha l, $p_{+c} = Y_{+c}/N$ proporção da população na coluna c, $Y_{l+} = \sum_{c=1}^{C} Y_{lc}$ total na linha l e $Y_{+c} = \sum_{l=1}^{L} Y_{lc}$ total na coluna c. Os somatórios das linhas e colunas são iguais a 1, $\sum_{c=1}^{C}\sum_{l=1}^{L} p_{lc} = 1$.

A proporção estimada de $\hat{p}_{lc}$ é calculada como a razão entre o total estimado para a casela (l,c) da tabela e o total geral estimado

$$\hat{p}_{lc} = \hat{Y}_{lc}/\hat{N} \tag{5.5}$$

De acordo com 5.2 e 5.3, essa expressão é representada por

$$\hat{p}_{lc} = \left(\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in S_{hi}} \delta_{hij}(l,c)w_{hij}\right) / \left(\sum_{h=1}^{H}\sum_{i\in S_h}\sum_{j\in S_{hi}} w_{hij}\right) \tag{5.6}$$

$\hat{p}_{lc}$ é a proporção estimada do indivíduo j do conglomerado i no estrato h que pertence a casela (l,c), sua variância estimada é representada por

$$\widehat{\text{var}}(\hat{p}_{lc}) = \sum_{h=1}^{H} \widehat{\text{Var}}_h(\hat{p}_{lc}) \tag{5.7}$$

com

$$\widehat{\text{var}}_h(\hat{p}_{lc}) = \sum_{h=1}^{H}\left\{\frac{S_h(1-f_h)}{S_h-1}\left[\sum_{i\in S_h}\sum_{j\in S_{hi}}\frac{(\delta_{hij}(l,c)-\hat{p}_{lc})}{\hat{N}}w_{hij} - \bar{\delta}_{hi}(l,c)\right]^2\right\}$$

e

$$\bar{\delta}_{hi}(l,c) = \sum_{i\in S_h}\sum_{j\in S_{hi}}\frac{[(\delta_{hij}(l,c)-\hat{p}_{lc})w_{hij}]/\hat{N}}{S_h}.$$

5.1 TESTE DE RAO-SCOTT

O teste Rao-Scott é um teste que ajusta a estatística qui-quadrado para o teste de independência e a qui-quadrado para o teste de homogeneidade, tal ajuste é incorporado para considerar o plano amostral adotado. E assim, obter a distribuição assintótica de referência, que é a qui-quidrado com $(L-1)(C-1)$ graus de liberdade. Nesta subseção é abordado o teste de Rao-Scott para hipótese nula de independência.

A hipótese nula de independência em termos de proporção é representada por

$$H_0 : p_{lc} = p_{l+}p_{+c} \quad l = 1, \ldots, L-1, \quad c = 1, \ldots, C-1. \tag{5.8}$$

A estatística qui-quadrado de Rao-Scott de primeira ordem depende apenas dos efeitos do plano amostral de proporção marginal (SAS, 2016), e é dada por

$$Q_{RS1} = X^2(I)/D \tag{5.9}$$

onde $X^2(I)$ é a estatística qui-quadrado (4.1) para o teste de independência e D é a correção do plano amostral a partir das estimativas de proporção, que tem a seguinte forma

$$D = \left(\sum_{l=1}^{L} \sum_{c=1}^{C} (1-\hat{p}_{lc}) \mathrm{Deff}(\hat{p}_{lc}) - \sum_{l=1}^{L} (1-\hat{p}_{l+}) \mathrm{Deff}(\hat{p}_{l+}) \right.$$
$$\left. - \sum_{c=1}^{C} (1-\hat{p}_{+c}) \mathrm{Deff}(\hat{p}_{+c}) \right) / (L-1)(C-1)$$

Deff é o efeito do plano amostral e é representado por

$$\begin{aligned} \mathrm{Deff}(\hat{p}_{lc}) &= \widehat{\mathrm{Var}}(\hat{p}_{lc}) / \widehat{\mathrm{Var}}_{aas}(\hat{p}_{lc}) \\ &= \widehat{\mathrm{Var}}(\hat{p}_{lc}) / (1-f)\hat{p}_{lc}(1-\hat{p}_{lc})/(n-1) \end{aligned}$$

em que $\hat{p}_{lc}$ a estimativa da proporção na casela (l, c), $\widehat{\mathrm{Var}}(\hat{p}_{lc})$ é a variância estimada de $\hat{p}_{lc}$ no plano amostral utilizado no estudo, $\widehat{\mathrm{Var}}_{aas}(\hat{p}_{lc})$ é a variância estimada de $\hat{p}_{lc}$ na amostra aleatória simples, f é a fração de amostragem global e n é o tamanho da amostra.

Sob a hipótese nula de nenhuma associação, a estatística qui-quadrado de Rao-Scott de primeira ordem segue aproximadamente uma distribuição qui-quadrado com $(L-1)(C-1)$ graus de liberdade. Uma melhor aproximação pode ser obtida pela estatística F

$$F_1 = Q_{RS1}/(L-1)(C-1)$$

que tem uma distribuição F com $(L-1)(C-1)$ e $\kappa(L-1)(C-1)$ graus de liberdade sob a hipótese nula (THOMAS e RAO (1984), THOMAS e RAO (1987)). O valor de κ é os graus

de liberdade do estimador de variância. O cálculo dos graus de liberdade depende do plano amostral e do método de estimativa de variância. Ainda existe a correção de Rao-Scott de segunda ordem para mais informações consulte (RAO; SCOTT, 1979), (RAO; SCOTT, 1981) e (RAO; THOMAS, 1989).

5.2 TESTE DE WALD

O teste de Wald é um teste em que a estatística investiga a hipótese nula de independência e de homogeneidade em tabelas de contingência de dupla entrada, levando em consideração o plano amostral da pesquisa. A seguir é mostrado o teste de Wald para hipótese de independência que é baseado na diferença entre as frequências observada e esperada na casela.

Sob a hipótese nula de independência das variáveis de linha e coluna, as frequências de células esperadas são calculadas como

$$\mathbb{E}_{lc} = \hat{Y}_{l+}\hat{Y}_{+c}/\hat{N}$$

em que, $\hat{Y}_{l+}$ total estimado da linha l, $\hat{Y}_{+c}$ total estimado da coluna c e $\hat{N}$ total estimado da população N. A hipótese nula (20) pode ser ser representada em termos de frequências ponderadas da população iguais às frequências esperadas (SAS, 2016), da seguinte forma

$$H_0 : G_{lc} = \hat{Y}_{lc} - \mathbb{E}_{lc} = 0 \quad \text{para todo} \quad l = 1, \ldots, L \quad \text{e} \quad c = 1, \ldots, C. \tag{5.10}$$

A estatística de Wald é calculada pela expressão

$$Q_W(I) = \hat{\mathbf{G}}'(\mathbf{H}\hat{\mathbf{V}}(\hat{\mathbf{Y}})\mathbf{H}')^{-1}\hat{\mathbf{G}}$$

no qual, $\hat{\mathbf{G}}$ é uma matriz de $(L-1)(C-1)$ diferenças entre as frequências ponderadas observadas e esperadas $(\hat{Y}_{lc} - \mathbb{E}_{lc})$, e $(\mathbf{H}\hat{\mathbf{V}}(\hat{\mathbf{Y}})\mathbf{H})'$ estima a variância de $\hat{\mathbf{G}}$.

$\hat{\mathbf{V}}(\hat{\mathbf{Y}})$ é a matriz de covariância das estimativas de frequência $\hat{Y}_{lc}$ da casela (l, c), $\mathbf{H}$ é a matriz $(L-1)(C-1)$ de RC que contém as derivadas parciais dos elementos de $\hat{\mathbf{G}}$ com respeito aos elementos de $\hat{\mathbf{Y}}$. Os elementos de $\mathbf{H}$ são calculados da seguinte forma, onde a denota uma linha diferente da linha l, e b denota uma coluna diferente da coluna c.

$$
\begin{aligned}
\frac{\partial \hat{G}_{lc}}{\partial \hat{Y}_{lc}} &= 1 - (\hat{Y}_{l+} + \hat{Y}_{+c} - \hat{Y}_{+c}\hat{Y}_{l+}/\hat{N})/\hat{N} \\
\frac{\partial \hat{G}_{lc}}{\partial \hat{Y}_{ac}} &= -(\hat{Y}_{l+} - \hat{Y}_{l+}\hat{Y}_{+c}/\hat{N})/\hat{N} \\
\frac{\partial \hat{G}_{lc}}{\partial \hat{Y}_{lb}} &= -(\hat{Y}_{+c} - \hat{Y}_{l+}\hat{Y}_{+c}/\hat{N})\hat{N} \\
\frac{\partial \hat{G}_{lc}}{\partial \hat{Y}_{ab}} &= \hat{Y}_{l+}\hat{Y}_{+c}/\hat{N}^2
\end{aligned}
$$

Sob a hipótese nula de independência, a estatística $Q_W(I)$ segue aproximadamente uma distribuição qui-quadrado com $(L-1)(C-1)$ graus de liberdade para grandes amostras.

É possivel fazer correções na estatística de Wald, utilizando a estatística F de Wald $ajustada$, essa usada para tabelas maiores que 2×2 (SAS, 2016), e a estatística F de Wald.

A estatística F de Wald $ajustada$ é represendada por

$$F_{Wajustada} = Q_W(I)\frac{f-(L-1)(C-1)-1}{f(L-1)(C-1)}$$

no qual, $f = i - h$, i número total de conglomerados e h de estratos na amostra. A estatística $F_{Wajustada}$ possui distribuição assintótica F com $(L-1)(C-1)$ e $f-(L-1)(C-1)-1$ graus de liberdade.

E a estatística F de Wald é dada por

$$F_W = \frac{Q_W(I)}{(L-1)(C-1)}$$

que tem distribuição assintótica F com $(L-1)(C-1)$ e f graus de liberdade.

6 APLICAÇÃO A DADOS DE COVID-19

Os dados empregados, nesta dissertação, foram os da Pesquisa Sorológica Continuar Cuidando, realizada pelo Governo do Estado da Paraíba, liderada pela Secretaria de Estado de Saúde da Paraíba, com apoio do Observatório de Síndromes Respiratórias da Universidade Federal da Paraíba (UFPB). A Sociedade para o Desenvolvimeto da Pesquisa Científica (SCIENCE) definiu o plano de amostragem, realizou a coleta dos questionários, processou e analisou os dados. E o Laboratório Central de Saúde Pública - Paraíba (LACEN-PB) em João Pessoa foi responsável pela análise do material colhido para o teste RT-PCR. A pesquisa teve como objetivo verificar a situação epidemiológica do estado da Paraíba frente à COVID-19 (SCIENCE, 2022).

O plano amostral adotado foi o estratificado e conglomerado em dois estágios, um plano complexo, realizado por Pedro Luis do Nascimento Silva, Mauricio Teixeira Leite de Vasconcellos, Hemilio Fernandes Campos Coelho e Cristiano Ferraz, descrito no relatório técnico Plano Amostral da Pesquisa PB-COVID19. No primeiro estágio do estudo efetuou a seleção dos setores censitários em cada estrato e utilizou o método de amostragem com probabilidades proporcionais ao tamanho (PPT) pelo método de Pareto ROSÉN (2000). Já no segundo estágio realizou o sorteio de domicílios em cada setor censitário e usou o método de amostragem Binomial ou de Bernoulli (SÄRNDAL; SWENSSON; WRETMAN, 1992), programado em dispositivos móveis de coleta.

Os estratos foram inicialmente definidos em 4 domínios (João Pessoa, Macro 1 sem João Pessoa, Macro 2 e Macro 3), que corresponde à estratificação natural da amostra. A estratificação foi concluída subdividindo os estratos naturais em estratos mais finos, buscando assegurar um espalhamento da amostra no território do estado, totalizando em 82 estratos no estudo. Os conglomerados da pesquisa são os domicílios e os setores censitários, com um total de 3.192 conglomerados.

O conjunto de dados reais foi obtido através de um questionário, com perguntas que abrangem o perfil sociodemográfico e de saúde das famílias, e ainda, os resultados do teste rápido (Medtest Coronavirus IgG/IgM - sangue) e do teste RT-PCR (com material coletado por swab na orofaringe), com um total de 10.872 pessoas (observações). O teste rápido de sangue que analisa o IgM, que identifica se possuem os anticorpos IgM, revelando infecção ativa, ou seja, a pessoa está infectada e transmitindo o vírus e/ou IgG designa que já ocorreu

infecção no passado e que a pessoa já está curada do COVID-19. E o teste RT-PCR constata se indivíduos têm o vírus causador do COVID-19 para resultado positivo (reagente) e negativo (não reagente). Os dados foram coletados nos meses de novembro e dezembro de 2020. As variáveis deste estudo foram sexo, resultado dos testes IgM e IgG sorológico, IgM rápido, RT-PCR e uso de máscara.

Utilizou-se o *Software* SAS OnDemand for Academics, uma versão gratuita destinada a estudantes e aprendizes independentes do SAS, para realizar as análises estatísticas. O procedimento empregado foi o **SURVEYFREQ**, que realiza análises de dados categóricos para dados de pesquisas complexas, como testes qui-quadrado (SAS, 2016). Com base nos dados da Pesquisa Sorológica Continuar Cuidando realizou-se testes para analisar a hipótese nula de independência, em todos considerou-se o nível de significância $\alpha = 0,05$. O objetivo é mostrar como não considerar o número de estratos e conglomerados podem afetar os resultados.

Na Tabela 12, a variável da linha é o resultado do IgM sorológico do último teste de COVID-19 e a variável da coluna é o sexo. O total de observação é de 416, visto que, as 10.445 são ausentes e 11 foram excluídas da análise devido ao baixo número de observações nas caselas correspondentes as categorias masculino e inconclusivo (1), feminino e inconclusivo (1), e, também, masculino e não sabe ou não recebeu (3). Nota-se, na Tabela 12, que das 96 pessoas que testaram positivo (reagente), 29 são do sexo masculino e 67 feminino, e, ainda, das 320 que testaram negativo (não-reagente), 132 pessoas são masculinas e 188 femininas. Percebe-se que a quantidade de pessoas do sexo feminino que realizaram o teste IgM sorológico é superior ao masculino. O interesse é verificar se as variáveis resultado do teste IgM sorológico e o sexo são independentes, no caso que, é considerado o plano amostral adotado na pesquisa e quando não é. As hipóteses a serem testadas são:

H_0 : O resultado do último teste IgM sorológico **não está associado** ao sexo da pessoa. (Independentes)

H_1 : O resultado do último teste IgM sorológico **está associado** ao sexo da pessoa. (Dependentes)

Tabela 12 – Frequência de pessoas, segundo o sexo e o resultado do teste IgM sorológico

Resultado do teste IgM sorológico por sexo			
Teste	Sexo		
IgM sorológico	Masculino	Feminino	Total
Positivo	29	67	96
Negativo	132	188	320
Total	161	255	416

Fonte: Elaborada pela autora (2022)

A hipótese nula é investigada pelo teste de Rao-Scott de primeira ordem, executa um estudo considerando o plano amostral adotado na pesquisa, com um total de 10.861 observações, pois, retirou-se 11, 82 estratos e 3.192 conglomerados, logo em seguida, é feito o teste sem levar em conta os estratos, depois, sem considerar os conglomerados e, por fim, sem manter ambos, seus resultados estão descritos na Tabela 13.

A estatística de teste de Rao-Scott de primeira ordem resulta em $7,79$, com 1 graus de liberdade (GL) e com um valor $p = 0,0053$, e a estatística F é $7,79$, com 1 GL no numerador e 232 GL denominador e o valor p é de $0,0057$, então como o valor $p < \alpha = 0,05$ (nível de significância) rejeita-se a hipótese nula de independência, sendo assim, existe uma associação entre o resultado do último teste IgM sorológico e o sexo. Nota-se que, quando retira os estrato da análise, os resultados são próximos dos obtidos considerado a plano original da pesquisa. Entretanto, quando a análise desconsidera os conglomerados, tanto a estatística de Rao-Scott $(6,44)$, com um valor $p = 0,0111$, quanto a $F\ (6,44)$, com um valor $p = 0,0116$, tem uma diferenciação do resultado obtido com o plano adotado na pesquisa. Acontece o mesmo ao remover os estratos e conglomerados, a estatística de Rao-Scott $(6,36)$ e $F\ (6,36)$ possuem valores $p\ 0,0116$ e $0,0120$, respectivamente. A decisão dos testes continua a mesma, rejeita a hipótese nula.

Tabela 13 – Resultado do teste de Rao-Scott, para as variáveis sexo e resultado do último teste IgM sorológico realizado, considerando o plano amostral, sem levar em conta os estratos, retirando os conglomerados e sem considerar estratos e conglomerados.

Resultado do teste de Rao-Scott				
	Plano amostral	Sem estratos	Sem conglomerados	Sem estratos e congl.
Rao-Scott	$7,79$	$7,61$	$6,44$	$6,36$
GL	1	1	1	1
valor p	$0,0053$	$0,0058$	$0,0111$	$0,0116$
Estatística F	$7,79$	$7,61$	$6,44$	$6,36$
GL (num.)	1	1	1	1
GL (den.)	232	303	344	415
valor p	$0,0057$	$0,0061$	$0,0116$	$0,0120$

Fonte: Elaborada pela autora (2022)

O teste de Wald para as variáveis da Tabela 12 (sem a categoria inclonclusivo e não sabe ou não recebeu) está representado na Tabela 14, o valor de sua estatística é $8,22$ quando é considerado o plano amostral da pesquisa, $7,65$ sem levar em conta os estratos, $7,06$ retirando os conglomerados e $6,69$ no momento em que é tirado os estratos e conglomerados. Nota-se que, em todas as análise a decisão do teste é rejeitar a hipótese de independência das variáveis sexo e resultado do último teste IgM, pois, os valores p são menores que o nível de significância $\alpha = 0,05$. Mas, são diferentes quando é considerado o plano amostral da pesquisa $(0,0045)$, sem levar em conta os estratos $(0,0060)$, retirando os conglomerados $(0,0082)$ e quando é tirado os estratos e conglomerados $(0,0100)$.

Tabela 14 – Resultado do teste de Wald, para as variáveis sexo e resultado do último teste IgM sorológico realizado, considerando o plano amostral, sem levar em conta os estratos, retirando os conglomerados e sem considerar estratos e conglomerados.

Resultado do teste de Wald				
	Plano amostral	Sem estratos	Sem conglomerados	Sem estratos e congl.
Wald	$8,22$	$7,65$	$7,06$	$6,69$
Estatística F	$8,22$	$7,65$	$7,06$	$6,69$
GL (num.)	1	1	1	1
GL (den.)	232	303	344	415
valor p	$0,0045$	$0,0060$	$0,0082$	$0,0100$

Fonte: Elaborada pela autora (2022)

A Tabela 15 possui um total de 265 observações, visto que, as 10.604 são ausentes e 3 foram excluídas da análise devido ao baixo número de observações nas caselas correspondentes as categorias masculino e inconclusivo (1), e, feminino e inconcluisovo(2). Distribuídas de acordo com o resultado de algum teste RT-PCR realizado (positivo, negativo e não sabe ou não recebeu) pra cada sexo (masculino e feminino), sendo que, 109 é a quantidade de masculino e 155 feminino. Foi realizado teste pra averiguar se as variáveis resultado de algum teste RT-PCR realizado e sexo são independentes, mantendo o plano amostral da pesquisa, retirando estratos, sem considerar conglomerados e sem ambos. As hipóteses a serem examinadas são:

H_0 : O resultado de algum teste RT-PCR realizado **não está associado** ao sexo da pessoa. (Independentes)

H_1 : O resultado de algum teste RT-PCR realizado **está associado** ao sexo da pessoa. (Dependentes)

Tabela 15 – Frequência de pessoas, segundo o sexo e o resultado de algum teste RT-PCR realizado.

Resultado do teste RT-PCR por sexo			
Teste	Sexo		
RT-PCR	Masculino	Feminino	Total
Positivo	35	37	72
Negativo	45	71	116
Não sabe ou não recebeu	29	48	77
Total	109	156	265

Fonte: Elaborada pela autora (2022)

O intuito de analisar essas hipóteses é mostrar que se não considerar o plano amostral em que os dados foram obtidos a decisão do teste pode chegar a conclusões errôneas. Para isso, realizou o teste de Rao-Scott de primeira ordem com o plano original, sem ele, e ainda, considerando-o parcialmente, os resultados estão expostos na Tabela 16.

Quando é levado em conta o plano amostral estabelecido na pesquisa a estatistica de Rao-Scott de primeira ordem é definida por $5,23$ com 2 graus de liberdade e valor p igual a $0,0730$, a estatística $F = 2,61$ com 2 graus de liberdade no numerador e 252 no denominador e valor $p = 0,0750$. Dessa forma, não rejeita a hipótese de não associação (independência), logo o resultado de algum teste RT-PCR realizado não tem relação com o sexo da pessoa. Ao comparar esse resultado com o obtido sem os estratos, que possui estatística de Rao-Scott $= 4,56$ com 2 GL e valor $p = 0,1020$ e $F = 2,28$ com graus de liberdade igual a 2 no numerador e 376 no denominador e valor $p = 0,1034$, permanece não rejeitando a hipótese nula, mas as estatística de Rao-Scott e F são diferentes em cada análise, acontece o mesmo com as verificações sem conglomerados e sem ambos.

Tabela 16 – Resultado do teste de Rao-Scott, para as variáveis sexo e resultado de algum teste RT-PCR realizado, considerando o plano amostral, considerando o plano amostral, sem levar em conta os estratos, retirando os conglomerados e sem considerar estratos e conglomerados.

Resultado do teste de Rao-Scott				
	Plano amostral	Sem estratos	Sem conglomerados	Sem estratos e congl.
Rao-Scott	$5,23$	$4,56$	$4,46$	$4,17$
GL	2	2	2	2
valor p	$0,0730$	$0,1020$	$0,1073$	$0,1240$
Estatística F	$2,61$	$2,28$	$2,23$	$2,08$
GL (num.)	2	2	2	2
GL (den.)	252	376	404	528
valor p	$0,0750$	$0,1034$	$0,1087$	$0,1251$

Fonte: Elaborada pela autora (2022)

O valor da estatística de Wald para verificar a independência das variáveis sexo e resultado de algum teste RT-PCR está representado na Tabela 17, sendo $4,22$ considerando o plano amostral da pesquisa, $3,35$ retirando os estratos, $3,61$ sem levar em conta os conglomerados e $3,09$ sem considerar os estratos e conglomerado. A decisão do teste, em todos os estudos, de acordo com a estatística F e F ajustada é não rejeitar a hipótese de independência, afirmando que as variáveis são independentes. Observa-se que os valores p resultantes tanto na estatística F $(0,1255; 0,1897; 0,1669; 0,2142)$, quanto na F *ajustada* $(0,1276; 0,1914; 0,1684; 0,2155)$ são muito diferentes em cada análise.

Tabela 17 – Resultado do teste de Wald, para as variáveis sexo e resultado de algum teste RT-PCR realizado, considerando o plano amostral, considerando o plano amostral, sem levar em conta os estratos, retirando os conglomerados e sem considerar estratos e conglomerados.

Resultado do teste de Wald				
	Plano amostral	Sem estratos	Sem conglomerados	Sem estratos e congl.
Wald	$4,22$	$3,35$	$3,61$	$3,09$
Estatística F	$2,11$	$1,67$	$1,80$	$1,54$
GL (num.)	2	2	2	2
GL (den.)	126	188	202	264
valor p	$0,1255$	$0,1897$	$0,1669$	$0,2142$
$F\ ajustada$	$2,09$	$1,66$	$1,79$	$1,54$
GL (num.)	2	2	2	2
GL (den.)	125	187	201	263
valor p	$0,1276$	$0,1914$	$0,1684$	$0,2155$

Fonte: Elaborada pela autora (2022)

Ademais, é executado estudo com a variável uso de máscara para maiores de 18 anos, com a finalidade de verificar sua relação com o resultado do último teste IgG sorológico realizado em planos amostrais que consideram sua formação e os que não consideram. É importante frisar que, ao realizar a análise com as 10.872 observações da pesquisa Continuar Cuidando a tabela de contingência, das variáveis de interesse, possue caselas com resultados iguais a zero, com isso, não é possível realizar os testes, optando em excluir observações e, então, prosseguir com as análises, de tal forma que, 13 observações são retiradas, referentes as categorias "inconclusivo (ou indeterminado)"e "não sabe ou ainda não recebeu"da variável resultado IgG sorológico, restando em 10.859, dessas 10.501 são ausentes. E o total de conglomerados é de 3.191.

A tabela 18, descreve a frequência relativa de 358 observações, que estão alocadas segundo o uso de máscara das pessoas maiores de 18 anos e o resultado do último teste IgG sorológico realizado. Constata-se que, há mais resultados negativos (278) para o teste IgG sorólógico que positivos (80), e, ainda, observa-se que, as pessoas que responderam quem usam máscara sempre 66 testaram positivo e 249 negativo. As hipóteses para investigar a associação dessas variáveis são descritas abaixo:

H_0 : O resultado do último teste IgG sorológico realizado **não está associado** ao uso de máscara das pessoas maiores de 18 anos. (Independentes)

H_1 : O resultado do último teste IgG sorológico realizado **está associado** ao uso de máscara das pessoas maiores de 18 anos. (Dependentes)

Tabela 18 – Frequência de pessoas, segundo o uso de máscara das pessoas maiores de 18 anos e o resultado do último teste IgG sorológico realizado.

Resultado do uso de máscara pelo teste IgG			
Uso de máscara	Resultado IgG		
	Positivo	Negativo	Total
Sempre	66	249	315
Às vezes ou nunca	14	29	43
Total	80	278	358

Fonte: Elaborada pela autora (2022)

A Tabela 19, apresenta o resultado do teste de Rao-Scott de primeira ordem e da estatística F para a análise da hipótese de não associação das variáveis. Lembrando que o nível de significância considerado é $\alpha = 0,05$. Nota-se que os estudos com o plano amostral adotado na pesquisa, a estatística de Rao-Scott é $3,45$ com um valor $p = 0,0631$, e o que não considera os estratos, possui estatística de Rao-Scott igual a $3,45$ com um valor $p = 0,0631$, o teste sem ponderar os conglomerados, a estatística de Rao-Scott é $3,31$ com um valor $p = 0,0686$, e sem levar em conta os estratos e conglomerados, a estatística de Rao-Scott é $3,31$ com um valor $p = 0,0687$, todas as análises, não rejeitam a hipótese nula, , pois os valores $p > \alpha = 0,05$, ou seja, as variáveis não estão associadas (independentes).

Tabela 19 – Resultado do teste de Rao-Scott, para as variáveis uso de máscara das pessoas maiores de 18 anos e o resultado do último teste IgG sorológico realizado, considerando o plano amostral, sem levar em conta os estratos, retirando os conglomerados e sem considerar estratos e conglomerados.

Resultado do teste de Rao-Scott				
	Plano amostral	Sem estratos	Sem conglomerados	Sem estratos e congl.
Rao-Scott	$3,45$	$3,45$	$3,31$	$3,31$
GL	1	1	1	1
valor p	$0,0631$	$0,0631$	$0,0686$	$0,0687$
Estatística F	$3,45$	$3,45$	$3,31$	$3,31$
GL (num.)	1	1	1	1
GL (den.)	209	277	289	357
valor p	$0,0645$	$0,0642$	$0,0696$	$0,0696$

Fonte: Elaborada pela autora (2022)

O valor da estatística de Wald é diferente em cada análise no estudo para analisar a independência das variáveis uso de máscara das pessoas maiores de 18 anos e resultado do último teste IgG, representado na Tabela 20. Nota-se que considerando o plano amostral adotado na pesquisa o valor é $2,49$, retirando os estratos é $2,48$, sem os conglomerados $2,50$ e sem levar em conta os estratos e conglomerados $2,47$. A decisão do teste, em todos os estudos, de acordo com a estatística F e é não rejeitar a hipótese de independência, afirmando que as variáveis são independentes. Observa-se que os valores p resultantes tanto na estatística F $(0,1150; 0,1159; 0,1159; 0,1163)$.

Tabela 20 – Resultado do teste de Wald, para as variáveis uso de máscara das pessoas maiores de 18 anos e o resultado do último teste IgG sorológico realizado, considerando o plano amostral, sem levar em conta os estratos, retirando os conglomerados e sem considerar estratos e conglomerados.

Resultado do teste de Wald				
	Plano amostral	Sem estratos	Sem conglomerados	Sem estratos e congl.
Wald	$2,49$	$2,48$	$2,50$	$2,47$
Estatística F	$2,49$	$2,48$	$2,50$	$2,47$
GL (num.)	1	1	1	1
GL (den.)	209	277	286	357
valor p	$0,1150$	$0,1159$	$0,1159$	$0,1163$

Fonte: Elaborada pela autora (2022)

Por fim, é realizado análises com as variáveis uso de máscara das pessoas maiores de 18 anos e resultado do teste IgM rápido, realizado no momento das visitas aos domicicílios no estado da Paraíba. Vale ressaltar que, no estudo para avaliar a associação das variáveis de interesse, houve, também, caselas com resultados iguais a zero, preferindo excluir as observações e, então, prosseguir com os testes, sendo que, da categoria "inconclusivo (ou indeterminado)"pertencentes a variável resultado do teste IgM rápido são excluídos 4 observações, restando 10.868, dessas 4.504 são ausentes, ficando 6.364 observações para as análises. E o total de conglomerados é de 3.192.

O total de observações distribuídas pela interseção das variáveis uso de máscara das pessoas maiores de 18 anos e resultado do teste IgM rápido, é de 6.364, representadas na Tabela 21. Verifica-se que, o total de teste que resultou em negativo para as pessoas que usam máscara sempre, às vezes e nunca é superior ao total de positivo, sendo que, de 5.272 pessoas que responderam que sempre usam máscara, 5.082 testou negativo e 190 positivo, de 957 que disseram que usa máscara às vezes, 929 testou negativo e 28 positivo. Além do mais, de 135 que nunca usam máscara, 127 testou negativo e 8 positivo. E ainda, nota-se que, 6.138 são negativos e 226 positivos. As hipóteses para explorar a associação dessas variáveis são dadas por:

H_0 : O resultado do teste IgM rápido **não está associado** ao uso de máscara das pessoas maiores de 18 anos. (Independentes)

H_1 : O resultado do teste IgM rápido **está associado** ao uso de máscara das pessoas

maiores de 18 anos. (Dependentes)

Tabela 21 – Frequência de pessoas, segundo o uso de máscara das pessoas maiores de 18 anos e o resultado do teste IgM rápido.

Resultado do teste IgM pelo uso de máscara			
Uso de máscara	Resultado IgM		
	Positivo	Negativo	Total
Sempre	190	5082	5272
Às vezes	28	929	957
Nunca	8	127	135
Total	226	6138	6364

Fonte: Elaborada pela autora (2022)

Para analisar a hipótese de não associação utilizou-se o teste de Rao-Scott de primeira ordem representado na Tabela 22, que mostra os resultados para o plano amostral original da pesquisa e sem considerá-lo, e ainda, levando-o em conta parcialmente. Percebe-se que, quando é analisado o teste no plano amostral adotado, a estatística de Rao-Scott é igual a $4,58$ com um valor $p = 0,1013$, e no sem considerar os estratos, a estatística de Rao-Scott é igual a $4,59$ com um valor $p = 0,1004$, a decisão é não rejeitar a hipótese nula, pois as os valores p são maiores que o nível de siginificância, $\alpha = 0,05$.

Portanto, as variáveis uso de máscara das pessoas maiores de 18 anos e resultado do teste IgM rápido são independentes. Todavia, quando é averiguado em plano que não considera os conglomerados, a estatística de Rao-Scott é igual a $6,11$ com um valor $p = 0,0471$, e sem levar em conta estratos e conglomerados, a estatística de Rao-Scott é igual a $6,10$ com um valor $p = 0,0471$, a decisão do teste muda, porque os valores p são menores que $\alpha = 0,05$.

Tabela 22 – Resultado do teste de Rao-Scott, para as variáveis uso de máscara das pessoas maiores de 18 anos e teste IgM rápido, considerando o plano amostral, sem levar em conta os estratos, retirando os conglomerados e sem considerar estratos e conglomerados.

Resultado do teste de Rao-Scott				
	Plano amostral	Sem estratos	Sem conglomerados	Sem estratos e congl.
Rao-Scott	$4,58$	$4,59$	$6,11$	$6,10$
GL	2	2	2	2
valor p	$0,1013$	$0,1004$	$0,0471$	$0,0471$
Estatística F	$2,29$	$2,29$	$3,05$	$3,05$
GL (num.)	2	2	2	2
GL (den.)	5698	5860	12564	12726
valor p	$0,1013$	$0,1005$	$0,0471$	$0,0472$

Fonte: Elaborada pela autora (2022)

Para investigar a hipótese de não associação do uso de máscara e resultado do teste IgM rápido utilizou-se, também, o teste de Wald, retratado na tabela 23, em que, o valor da estatística de Wald é igual no estudo considerando o plano amostral adotado na pesquisa e sem os estratos, $2,26$, sem levar em conta os conglomerados é $3,25$ e retirando estratos e conglomerados é de $3,24$. A estatística F correspondente, também, é igual no estudo considerando o plano amostral adotado na pesquisa e sem os estratos, $1,14$, com o mesmo valor p, $0,3173$, então, não rejeita a hipótese de independência, concluindo que as variáveis são indepententes. Nota que os valores p da análise das variáveis sem considerar os conglomerados, $0,1966$, e retirando estratos e conglomerados, $0,1977$, de acordo com a estatística F correspondente, são muito inferiores aos dos estudos do plano original e retirando os estratos, e continuam não rejeitando a hipótese de independência. As conclusões do teste, de acordo com a estatística F *ajustada* são as mesmas das obtidas na estatística F.

Tabela 23 – Resultado do teste de Wald, para as variáveis uso de máscara das pessoas maiores de 18 anos e teste IgM rápido, considerando o plano amostral, sem levar em conta os estratos, retirando os conglomerados e sem considerar estratos e conglomerados.

Resultado do teste de Wald				
	Plano amostral	Sem estratos	Sem conglomerados	Sem estratos e congl.
Wald	$2,29$	$2,29$	$3,25$	$3,24$
Estatística F	$1,14$	$1,14$	$1,62$	$1,62$
GL (num.)	2	2	2	2
GL (den.)	2849	2930	6282	6363
valor p	$0,3173$	$0,3173$	$0,1966$	$0,1977$
$F\ ajustada$	$1,14$	$1,14$	$1,62$	$1,62$
GL (num.)	2	2	2	2
GL (den.)	2848	2929	6281	6362
valor p	$0,3174$	$0,3175$	$0,1966$	$0,1977$

Fonte: Elaborada pela autora (2022)

Constata-se que os resultados obtidos no teste de Rao-Scott e de Wald são diferentes para a mesma análise. Segundo Lohr (2010), o teste de Wald é recomendado quando é utilizado uma estimativa apropriada da $(\hat{\mathbf{V}}(\hat{\mathbf{Y}}))$ matriz de covariância das estimativas de frequência $\hat{Y}_{lc}$, porque fornece uma estatística de teste assintoticamente válida. Mas, em tabelas maiores que 2×2 pode haver desvantagem, devido a instabilidade nas estimativas. O teste de Rao-Scott é indicado quando não é possível obter uma estimativa apropriada da $\hat{\mathbf{V}}(\hat{\mathbf{Y}})$. Na prática, o teste de Wald não é recomendado em amostras "grandes", que em amostra complexa se refere a um grande número de unidades primárias de amostragem, em razão que, a matriz de covariância estimada $(\hat{\mathbf{V}}(\hat{\mathbf{Y}}))$ será muito instável.

As análises demonstram que as estatísticas mudam no estudo, como também a decisão do teste pode ser alterada, quando é comparado ao resultado obtido considerando o plano amostral da pesquisa, acarretando em resultados errôneos. Nota-se, ainda, que há um impacto maior quando é retirado os conglomerados, e também, quando é retirado estratos e conglomerados no estudo. Recomenda-se sempre considerar o plano amostral adotado na pesquisa.

7 CONCLUSÕES

Os testes de associação como o de Rao-Scott e de Wald são os usados para analisar a relação das variáveis em tabelas de contingência com dados complexos, eles consideram o plano amostral adotado na pesquisa. Esses testes permitem estudar a independência, que verifica se o fator de linha e de coluna não estão associados, e a homogeneidade, que analisa se duas ou mais amostras independentes variam em distribuição em um único fator.

Na presente dissertação realizada, deseja-se mostrar os principais testes de associação para tabelas de contingência para dados amostrais complexos. E apresentar a importância de considerar o plano amostral que obteve os dados, através de ilustrações de análises de dados oriundos da pesquisa sorológica Continuar Cuidando, do estado da Paraíba, que teve como objetivo estimar a prevalência de casos de COVID-19 no estado.

O estudo para investigar a independência das variáveis sexo e resultado do último teste IgM, com o teste de Rao-Scott de primeira ordem, mostrou que existe associação entre elas, isto é, são dependentes, a um nível de significância de 5%, tanto na análise considerando o plano amostral da pesquisa quanto considerando-o parcialmente. Todavia, o valor da estatística em cada análise é diferente.

Quando é averiguado a independência das variáveis uso de máscara das pessoas maiores de 18 anos e o resultado do teste IgM, através do teste de Rao-Scott de primeira ordem a um nível de significância de 5%, considerando o plano amostral adotado na pesquisa e sem considerar os estratos, constata que não são associadas, ou seja, são independentes. No entanto, quando é analisado o estudo com o plano amostral retirando o conglomerados e retirando ambos, estabelece que são associadas, isto é, são dependentes.

No teste de Wald para analisar, também, a independência das variáveis, todos os estudos resultaram em afirmar que as variáveis não estão associadas. Salvo, a investigação do sexo com o resultado do último teste IgM sorológico, que concluiu que as variáveis são dependentes, isto é, estão associadas. Observa-se que os valores p das estatisticas F e F *ajustada* são diferentes nas análises.

Salienta-se ainda que, há diferença nos testes de Rao-Scott e de Wald, pois o teste de Rao-Scott é indicado quando não é possível obter uma estimativa apropriada da $(\hat{\mathbf{V}}(\hat{\mathbf{Y}}))$ matriz de covariância das estimativas de frequência $\hat{Y}_{lc}$ e o teste de Wald é recomendado quando é utilizado uma estimativa apropriada da $\hat{\mathbf{V}}(\hat{\mathbf{Y}})$, visto que, quando é adequada fornece uma

estatística de teste assintoticamente válida.

Portanto, conclui-se que em estudo com dados complexos dispostos em tabelas de contingência é necessário considerar a estratificação, conglomeração e/ou probabilidade proporcional a uma medida de tamanho, ou seja, as características do plano amostral que obteve os dados, por meio dos testes que os consideram. Para não ocasionar em resultados das estatísticas e conclusões errôneas.

REFERÊNCIAS

BUSSAB, W. d. O.; MORETTIN, P. A. *Estatística Básica*. 5. ed. [S.l.]: São Paulo: Saraiva, 2002.

GORDIS, L. *Epidemiologia*. [S.l.]: Thieme Revinter Publicações LTDA, 2017.

HORVITZ, D. G.; THOMPSON, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, Taylor & Francis, v. 47, n. 260, p. 663–685, 1952.

LOHR, S. L. *Sampling: Design and Analysis*. 2. ed. [S.l.]: Cengage Learning, 2010.

PAULINO, C. D.; SINGER, J. da M. *Análise de dados categorizados*. [S.l.]: Editora Blucher, 2006.

PEARSON, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor & Francis, v. 50, n. 302, p. 157–175, 1900.

PESSOA, D. G. C.; SILVA, P. L. N. *Análise de dados amostrais complexos*. [S.l.]: São Paulo: Associação Brasileira de Estatística, 1998.

RAO, J. N. K.; SCOTT, A. J. Chi-squared tests for analysis of categorical data from complex surveys. In: *Proceedings of the American Statistical Association, section on survey research methods*. [S.l.: s.n.], 1979. p. 58–66.

RAO, J. N. K.; SCOTT, A. J. The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American statistical association*, Taylor & Francis, v. 76, n. 374, p. 221–230, 1981.

RAO, J. N. K.; THOMAS, D. R. Chi-squared tests for contingency tables. *Analysis of complex surveys*, John Wiley & Sons, Inc., New York, p. 89–114, 1989.

ROSÉN, B. *A User's Guide to Pareto nps Sampling*. [S.l.]: Stockholm, Sweden: [s.n.], 2000.

SAS, I. I. SAS/STAT. 14.2 User's Guide. *Cary, NC: SAS Institute Inc*, 2016.

SCIENCE. *SCIENCE - Sociedade para o Desenvolvimento da Pesquisa Científica*. 2022. Disponível em: <https://science.org.br/research/pesquisa-continuar-cuidando-observatorio-da-covid-19/>. Acesso em: 17 maio 2022.

SINGER, M.; BULLED, N.; OSTRACH, B.; MENDENHALL, E. Syndemics and the biosocial conception of health. *The lancet*, Elsevier, v. 389, n. 10072, p. 941–950, 2017.

STOKES, M. E.; DAVIS, C. S.; KOCH, G. G. *Categorical data analysis using SAS*. 3. ed. [S.l.]: SAS institute, 2012.

SÄRNDAL, C. E.; SWENSSON, B.; WRETMAN, J. *Model Assisted Survey Sampling*. [S.l.]: New York: Springer-Verlag, 1992.

THOMAS, D. R.; RAO, J. N. K. A Monte Carlo study of exact levels of goodness-of-fit statistics under cluster sampling. In: *Proceedings of the American Statistical Association, section on survey research methods. Washington DC*. [S.l.: s.n.], 1984.

THOMAS, D. R.; RAO, J. N. K. Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, Taylor & Francis, v. 82, n. 398, p. 630–636, 1987.

APÊNDICE A – CÓDIGO SAS

```
/*************************************************************/
/* Programa para análise de dados amostrais complexos da     */
/* Pesquisa Sorológica Continuar Cuidando do estado da Paraíba */
/*-----------------------------------------------------------*/
/* Estimativas de prevalencia                                */
/* Analise de tabelas de contingencia sob modelo multinomial */
/*-----------------------------------------------------------*/
/*Abril 2022                                                 */
/*************************************************************/

/* Importacao dos dados */
FILENAME REFFILE '/home/u50384024/my_shared_file_links/cferraz/
Noemir/pbcovid_dat.sav';

PROC IMPORT DATAFILE=REFFILE
DBMS=SAV
OUT=WORK.IMPORT;
RUN;

/* Renomeacao de arquivo fonte */
data pb_covid19;
set import;
setor=cod_setor;
domicilio=a06_domicilio;
estratos=estrato_sel;
peso=peso_calibrado;
run;

/*Analise com as variaveis sexo e
e resultado do teste IgM sorologico */
```

```
/*excluindo a categoria inconclusivo (ou indeterminado)
e não sabe ou ainda não recebeu o resultado da variavel
resultado do teste igm rapido */

data Tabela12;
set pb_covid19;
if i06_igm_sorologico = 3 then delete;
if i06_igm_sorologico = 4 then delete;
run;

/*Tabela 12*/
proc freq data=Tabela12;
tables b02_sexo*i06_igm_sorologico;
run;

/*teste de Rao-Scott e Wald, considerar o plano
amostral da pesquisa, para verificar
a independencia das variaveis
uso mascara e resultado do teste IgM rapido*/

title 'Pesquisa Sorológica Continuar Cuidando';
proc surveyfreq  data=Tabela12;
tables b02_sexo*i06_igm_sorologico /row cl chisq
wchisq;
strata estratos;
cluster setor domicilio;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os estratos,
 para verificar a independencia das variaveis
 uso mascara e resultado do teste IgM rapido*/
```

```
title 'Pesquisa Sorológica Continuar Cuidando';
proc surveyfreq  data=Tabela12;
tables b02_sexo*i06_igm_sorologico /row cl chisq wchisq;
cluster setor domicilio;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os
conglomerados, para verificar a independencia das
variaveis usou mascara e resultado do teste IgM
rapido*/

title 'Pesquisa Sorológica Continuar Cuidando';
proc surveyfreq  data=Tabela12;
tables b02_sexo*i06_igm_sorologico /row cl chisq
wchisq;
strata estratos;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os estratos
e conglomerados, para verificar a independencia das
variaveis usou mascara e resultado do teste IgM rapido*/

title 'Pesquisa Sorológica Continuar Cuidando';
proc surveyfreq  data=Tabela12;
tables b02_sexo*i06_igm_sorologico /row cl chisq wchisq;
weight peso;
run;

/*********************************************************/
/*Analise com as variaveis sexo e resultado do teste rt-pcr*/
```

```
/*excluindo a categoria inconclusivo (ou indeterminado)
da variavel resultado do teste rt-pcr */

DATA Tabela15;
    SET pb_covid19;
    IF i09_resultado_rt_pcr = 3 THEN DELETE;
RUN;



/*Tabela 15*/
proc freq data=Tabela15;
tables b02_sexo*i09_resultado_rt_pcr;
run;

/*teste de Rao-Scott e Wald, considerando o plano
amostral da pesquisa, para verificar a independencia
das variaveis sexo e resultado do teste RT-PCR*/

PROC SURVEYFREQ DATA= Tabela15;
TABLES b02_sexo*i09_resultado_rt_pcr / row cl chisq wchisq;
strata estratos;
cluster setor domicilio;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os
estratos, para verificar a independencia das variaveis
sexo e resultado do teste RT-PCR*/

PROC SURVEYFREQ DATA= Tabela15;
TABLES b02_sexo*i09_resultado_rt_pcr / row cl chisq wchisq;
cluster setor domicilio;
```

```
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os
conglomerados,  para verificar a independencia das
variaveis sexo e resultado do teste RT-PCR*/

PROC SURVEYFREQ DATA= Tabela15;
TABLES b02_sexo*i09_resultado_rt_pcr / row cl chisq wchisq;
strata estratos;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os estratos
e conglomerado, para verificar a independencia das
variaveis sexo e resultado do teste RT-PCR*/

PROC SURVEYFREQ DATA= Tabela15;
TABLES b02_sexo*i09_resultado_rt_pcr / row cl chisq wchisq;
weight peso;
run;

/*********************************************************/
/*Analise com as variaveis uso de mascara e resultado do
teste IgG sorologico*/

/*excluindo a categoria inconclusivo(ou indeterminado)
e não sabe ou ainda não recebeu da variavel resultado
do teste IgG sorologico, e, também, não sabe ou não
respondeu da variavel uso de mascara */

data Tabela18;
set pb_covid19;
```

```
if i07_igg_sorologico=3 then delete;
if i07_igg_sorologico=4 then delete;
if f03_usou_mascara=4 then delete;
run;

/*tabela com os dados do arquivo pb_covid19*/
proc freq data=pb_covid19;
tables i07_igg_sorologico*f03_usou_mascara;
run;

/*tabela com as categorias acima retiradas*/
proc freq data=Tabela18;
tables i07_igg_sorologico*f03_usou_mascara;
run;

/*Juntando as categorias às vezes e nunca da
variável usou máscara*/

data Tabela18_colapsada;
set Tabela18;
if f03_usou_mascara=3 then f03_usou_mascara=2;
run;

/*Tabela 18 colapsada*/
proc freq data=Tabela18_colapsada;
tables i07_igg_sorologico*f03_usou_mascara;
run;

/*teste de Rao-Scott e Wald, considerando o plano
amostral da pesquisa, para verificar a independencia
das variaveis usou mascara e resultado do teste IgG
sorologico*/
```

```
proc surveyfreq data=Tabela18_colapsada;
tables f03_usou_mascara*i07_igg_sorologico /row cl chisq
wchisq;
strata estratos;
cluster setor domicilio;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os
estratos, para verificar a independencia
das variaveis usou mascara e resultado do teste IgG
sorologico*/

proc surveyfreq data=Tabela18_colapsada;
tables f03_usou_mascara*i07_igg_sorologico /row cl chisq
wchisq;
cluster setor domicilio;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os conglomerados,
para verificar a independencia
das variaveis usou mascara e resultado do teste IgG
sorologico*/

proc surveyfreq data=Tabela18_colapsada;
tables f03_usou_mascara*i07_igg_sorologico /row cl chisq
wchisq;
strata estratos;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerando os estratos
```

```
e conglomerados, para verificar a independencia
das variaveis usou mascara e resultado do teste IgG
sorologico*/

proc surveyfreq data=Tabela18_colapsada;
tables f03_usou_mascara*i07_igg_sorologico /row cl chisq
wchisq;
weight peso;
run;

/*************************************************************/

/*Analise com as variaveis usou mascara e
resultado do teste igm rapido*/

/*excluindo a categoria inconclusivo(ou indeterminado)
da variavel resultado do teste igm rapido */
DATA Tabela21;
    SET pb_covid19;
    IF j03_igm_teste_rapido = 3 THEN DELETE;
RUN;

/*tabela 21*/
proc freq data=Tabela21;
tables f03_usou_mascara*j03_igm_teste_rapido;
run;

/*teste de Rao-Scott e Wald, comsiderando o plano
amostral da pesquisa, para verificar a independencia
das variaveis usou mascara e resultado do teste igm rapido*/
title 'Pesquisa Sorológica Continuar Cuidando';
proc surveyfreq  data=Tabela21;
tables f03_usou_mascara*j03_igm_teste_rapido /row cl chisq wchisq;
```

```
strata estratos;
cluster setor domicilio;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os
estratos, para verificar a independencia das
variaveis usou mascara e resultado do teste igm
rapido*/
title 'Pesquisa Sorológica Continuar Cuidando';
proc surveyfreq  data=Tabela21;
tables f03_usou_mascara*j03_igm_teste_rapido /row cl chisq wchisq;
cluster setor domicilio;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os
conglomerados, para verificar a independencia das
variaveis usou mascara e resultado do teste igm
rapido*/
title 'Pesquisa Sorológica Continuar Cuidando';
proc surveyfreq  data=Tabela21;
tables f03_usou_mascara*j03_igm_teste_rapido /row cl chisq wchisq;
strata estratos;
weight peso;
run;

/*teste de Rao-Scott e Wald, sem considerar os estratos
e conglomerados, para verificar a independencia das
variaveis usou mascara e resultado do teste igm
rapido*/
title 'Pesquisa Sorológica Continuar Cuidando';
proc surveyfreq  data=Tabela21;
```

```
tables f03_usou_mascara*j03_igm_teste_rapido /row cl chisq wchisq;
strata estratos;
weight peso;
run;
```