

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

RAFAEL D'ANGELO REIS PEREIRA DA SILVA

AMOSTRAGEM PARA MODOS MISTOS DE PESQUISA: WEB E TELEFONE

JUIZ DE FORA

2024

RAFAEL D'ANGELO REIS PEREIRA DA SILVA

AMOSTRAGEM PARA MODOS MISTOS DE PESQUISA: WEB E TELEFONE

Monografia apresentada pelo(a) acadêmico(a) Rafael d'Angelo Reis Pereira da Silva ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para obtenção do título de Bacharel em Estatística.

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

Co-Orientador: Prof. Lupércio França Bessegato, Ph.D.

Juiz de Fora

2024

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

Silva, Rafael d'Angelo Reis Pereira.

Amostragem Para Modos Mistos de Pesquisa : Web e Telefone /
Rafael d'Angelo Reis Pereira Silva. -- 2024.
52 f. : il.

Orientador: Marcel de Toledo Vieira

Coorientador: Lupércio França Bessegato

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2024.

1. Modos Mistos. 2. Pesquisa online/web. 3. Pesquisa por telefone.
4. Amostragem Complexa. I. Vieira, Marcel de Toledo, orient. II.
Bessegato, Lupércio França, coorient. III. Título.

RAFAEL D'ANGELO REIS PEREIRA DA SILVA

AMOSTRAGEM PARA MODOS MISTOS DE PESQUISA: WEB E TELEFONE

Monografia apresentada pelo(a) acadêmico(a) Rafael d'Angelo Reis Pereira da Silva ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para obtenção do título de Bacharel em Estatística.

Aprovado em 27 de setembro de 2024

BANCA EXAMINADORA

Prof. Dr. Marcel de Toledo Vieira - Orientador
Universidade Federal de Juiz de Fora

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

Prof. Dr. Ronaldo Rocha Bastos
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

Em primeiro lugar, agradeço aos professores que me ajudaram a chegar onde estou, e que me ajudaram a montar os degraus para o meu futuro. Um agradecimento especial ao professor Lupércio, que me acompanhou desde meu primeiro período até o presente, pela sua paciência e dedicação aos alunos e especialmente à mim, e ao professor Marcel, no qual o apoio, aconselhamento e confiança foram essenciais para a criação deste trabalho, além de me inspirar a seguir esta linha de pesquisa com suas aulas de amostragem.

Agradeço à minha parceira Poliana Carvalho, por todo amor e carinho necessário para enfrentar as partes difíceis da vida. Mesmo estando em momentos tão difíceis quanto eu, sempre se ofereceu para ajudar e ser um pilar de esperança e confiança de que tudo daria certo no final. Sem ela este projeto não poderia ter sido feito.

Agradeço aos meus amigos, Yan, Pietro, André, Caio e Luan, pelas risadas que compartilhamos e pelos tempos de lazer que me permitiram completar esta árdua jornada. Inspiro-me em vocês a pessoa que desejo ser um dia.

Aos meus companheiros de estudos, Lucas, Deiverson, Maria, Matheus e Natália, pela considerável ajuda durante todos os anos de curso desde que nos conhecemos. Tenho certeza de que sozinho eu nunca teria chegado tão longe.

Agradeço aos meus pais por todo o apoio e suporte durante meus anos de estudo, que me permitiram me tornar a melhor versão de mim mesmo. Também à minha irmã Júlia, que durante toda minha vida esteve lá por mim, desde quando eu ainda não sabia apreciar seus ensinamentos e delicadezas.

Sem cada uma destas pessoas eu não seria a pessoa que sou hoje. Espero um dia poder retribuir o apoio que recebi de forma exponencial à todos.

*“Você segue os passos daqueles que vieram antes de você,
e seu caminho guia aqueles que o seguirão mais tarde.”
(Outer Wilds, 2019, jogo eletrônico)*

RESUMO

A coleta de dados desempenha um papel fundamental na pesquisa, moldando a qualidade e a maneira que os resultados de um estudo são apresentados. Uma das definições mais importantes em uma pesquisa, portanto, é a escolha de qual método de amostragem utilizar para a coleta dos dados, já que todas as opções apresentam suas próprias desvantagens e benefícios. Tendo isto em mente, o objetivo deste estudo foi uma revisão sobre a precisão e o desempenho de métodos mistos de amostragem, com foco em entrevistas online e por telefone, comparando seus vieses em aplicação e resposta. Foi utilizada uma gama de parâmetros para julgar os dois métodos; entre eles, analisamos as taxas de resposta em diferentes casos ao longo dos anos, os custos envolvidos em ambos métodos, tempo gasto para finalização da coleta, e combinações de diferentes métodos simultaneamente. Também realizamos um experimento envolvendo duas amostras de escolas da Zona da Mata Mineira, a partir das quais comparamos a qualidade das respostas de questionários aplicados por e-mail e por telefone em relação aos dados do Censo Educacional disponibilizado pelo INEP.

Palavras-chave: modos mistos, amostragem complexa, pesquisa online/web, pesquisa por telefone, efeito de modo.

ABSTRACT

Data collection plays a fundamental role in research, shaping the quality and way in which the results of a study are presented. One of the most important decisions in research, therefore, is the choice of which sampling method to use for data collection, since all options have their own disadvantages and benefits. With this in mind, the aim of this study was to review the accuracy and performance of mixed sampling methods, focusing on online and telephone interviews, comparing their bias in application and response. A range of parameters were used to judge the two methods; including the analysis of response rates in different cases over the years, the costs involved in both methods, the time spent to complete the collection, and combinations of different methods simultaneously. We also conducted an experiment involving two samples of schools in the Zona da Mata Mineira, from which we compared the quality of responses to questionnaires applied by e-mail and by telephone in relation to data from the Educational Census made available by INEP.

Keywords: mixed-mode, complex sampling, online/web, telephone, mode effect.

LISTA DE ILUSTRAÇÕES

Figura 1	– Fluxograma do estudo, com estratos e exclusões da população.....	31
Gráfico 1	– Número de Escolas da População por Tipo de Dependência	32
Gráfico 2	– Intervalos de Confiança da coleta por E-mail - Médias.....	39
Gráfico 3	– Intervalos de Confiança da coleta por Telefone - Médias.....	39
Gráfico 4	– Intervalos de Confiança da coleta por E-mail - Proporções.....	40
Gráfico 5	– Intervalos de Confiança da coleta por Telefone - Proporções.....	40

LISTA DE TABELAS

Tabela 1	– Resultados das médias para a coleta por e-mail	35
Tabela 2	– Resultados das proporções para a coleta por e-mail	36
Tabela 3	– Resultados das médias para a coleta por telefone	36
Tabela 4	– Resultados das proporções para a coleta por telefone	37
Tabela 5	– Resultados intervalos de confiança para a coleta por e-mail.....	37
Tabela 6	– Resultados intervalos de confiança para a coleta por telefone	38

SUMÁRIO

1. INTRODUÇÃO.....	11
2. USO DE MODOS MISTOS.....	12
2.1 MEIOS DE CONTATO.....	13
2.2 VIÉS DOS MODOS.....	15
2.3 DESIGN DO QUESTIONÁRIOS.....	16
3. ANÁLISE COMPARATIVA DE QUESTIONÁRIOS POR WEB E TELEFONE.....	18
3.1 A DECRESCENTE EFICÁCIA DE PESQUISAS TELEFÔNICAS.....	18
3.2 VANTAGENS ESTRATÉGICAS DE PESQUISAS POR TELEFONE.....	19
3.3 SMARTPHONES E A FALTA DE CONFIANÇA DO PÚBLICO, DESAFIOS DE PESQUISAS ONLINE.....	21
3.4 MOTIVOS PARA O OTIMISMO.....	22
4. REFERENCIAL TEÓRICO.....	24
4.1 AMOSTRAGEM ESTRATIFICADA.....	24
4.2 INTERVALOS DE CONFIANÇA.....	26
5. MÉTODOS.....	27
5.1 CENSO ESCOLAR DO INEP.....	27
5.2 VARIÁVEIS.....	28
5.3 COLETA E ANÁLISE DOS DADOS.....	29
5.4 CÁLCULOS.....	34
5.5 RESULTADOS.....	35
6. CONCLUSÕES E CONSIDERAÇÕES FINAIS.....	41
7. LIMITAÇÕES DO ESTUDO.....	42
REFERÊNCIAS BIBLIOGRÁFICAS.....	44
ANEXO A - CÓDIGOS PARA OS DADOS DA POPULAÇÃO.....	46
ANEXO B - CÓDIGOS PARA OS DADOS DA COLETA POR E-MAIL.....	49
ANEXO C - CÓDIGOS PARA OS DADOS DA COLETA POR TELEFONE.....	52

1. INTRODUÇÃO

O aumento das pesquisas online nas últimas duas décadas despertou um forte interesse em modos mistos de pesquisas (também conhecido como *mixed-mode*). Os questionários da *Web* são baratos e rápidos, mas possuem taxas de resposta mais baixas do que os modos tradicionais de questionário em papel, tanto aqueles baseados em entrevistas pessoais ou enviados pelo correio. Combinar pesquisas na *web* com outro modo de pesquisa, como entrevistas por telefone, é, portanto, uma maneira de aumentar as baixas taxas de resposta na *web*.

A crise da COVID-19 entre 2020 e 2021 deu um forte impulso à reflexão sobre pesquisas online e de modos mistos, uma vez que as entrevistas presenciais não eram possíveis durante a pandemia. A nova situação deu origem a um novo inventário aos Institutos de Estatística Nacionais (NSIs) europeus sobre as suas experiências durante a crise da COVID (Beck et al., 2022, *apud* Schouten et al., 2023). O inventário mostrou que os modos mistos se expandiram rapidamente, embora as entrevistas presenciais tenham sido substituídas principalmente por entrevistas por telefone, enquanto o nível de pesquisas pela *web* permaneceu estável. Pareceu muito mais complexo introduzir novos modos do que expandir o uso de modos que já eram familiares.

Apesar da mistura de modos permitir maior flexibilidade em circunstâncias imprevistas, nem todos pesquisadores visam os modos mistos como sua forma preferida de coleta. A flexibilidade vem acompanhada de complexidade adicional, tanto na elaboração de questionários, como na logística do trabalho de campo, no ajuste e na análise. Por outro lado, os custos crescentes do trabalho de campo e os problemas de cobertura em pesquisas de modo único forçam pesquisadores a adotar soluções de modos mistos e a incluir a *web* como opção. Se empresas e países consideram a continuação de modos mistos de pesquisa, é principalmente por razões de custo-eficiência (Beck et al., 2022, *apud* Schouten et al., 2023).

Nota-se que, pesquisas online autoadministradas foram propostas como um simples substituto para algumas das pesquisas que eram realizadas presencialmente. No entanto, uma pesquisa existente não é facilmente alterada desta forma. No lugar, uma mudança de procedimento é mais provável de ser aplicada a estudos que possuam requisitos de pesquisa simples e flexíveis.

O uso de pesquisas online como primeira ou segunda opção para a coleta é comumente usado em estudos que utilizam modos mistos. Isso permite que a amostra se torne

melhor distribuída em diferentes aspectos como a idade dos respondentes, níveis de escolaridade e renda ou afinidade com a tecnologia.

Um ponto relevante a ser considerado, entretanto, é a formulação do questionário para pesquisas que utilizem a coleta por *web* em conjunto a outros modos mais tradicionais. Desktops, notebooks, tablets, celulares e até mesmo aparelhos televisivos podem ter acesso à internet. Devido a isso, são fortemente discutidas as influências de modos mistos em erros de amostragem, custos e resultados gerados.

Dentro deste contexto, o objetivo deste estudo, é analisar historicamente o uso de *web* e telefones para a coleta de dados ao longo dos anos, comparando assim as vantagens e desvantagens de ambos os métodos. Com este conhecimento, utilizar os métodos de coleta de dados por *web* e por telefone em uma mesma população conhecida, utilizando os microdados do Censo Escolar do ano de 2023 como cadastro base para seleção das amostras, disponibilizados pelo INEP, com objetivo de compreender o funcionamento dos métodos de coleta, assim como seus resultados e apontar as características individuais de ambos os modos.

Na Seção 2 deste estudo, é discutida a utilização de modos mistos para coleta de dados, assim como seus benefícios e desafios encontrados em sua utilização. A seção 3 expande as particularidades de ambos os modos discutidos neste estudo, demonstrando as propriedades e exigências que devem ser aplicadas ao utilizá-los em uma amostra. Uma breve análise sobre os métodos de inferência de erros de medida é realizada na seção 4. Durante a sessão 5 descrevemos nossa pesquisa e amostras realizadas a partir do Censo Escolar do INEP. Por fim, o estudo se conclui na seção 6, com a discussão dos resultados obtidos.

2. USO DE MODOS MISTOS

Em uma pesquisa do tipo “*survey*”, há vários aspectos a considerar: que modos de coleta escolher, em que sequência aplicar os questionários e que estratégia de comunicação aplicar. Cada modo possui suas próprias vantagens e desvantagens, desde entrevistas cara a cara, por telefone, e-mail ou por cartas. Implícitas estão as considerações de custos, a prevenção da não resposta e do viés de seleção, e a otimização da qualidade dos dados. Além disso, uma maneira efetiva de prevenir os erros de medição específicos de diferentes modos, é a utilização de modos mistos, isto é, a combinação de diferentes modos, como telefone e e-mail, em uma mesma pesquisa.

Schouten et al. (2023) distinguem seis características dos modos de coleta: intimidade, interação, assistência, velocidade e ritmo, apresentação (oral ou visual) e tempo. Esses recursos impactam a experiência de realização da pesquisa e diferem em grande medida entre os modos. A maior diferença vem dos modos com a presença de um entrevistador e aqueles auto administrados, que mostram disparidade em todas as características. Além disso, os modos também podem introduzir comportamentos e estilos de resposta específicos. Por exemplo, os modos com entrevistadores, através da sua maior intimidade, interação e assistência, podem ajudar a evitar erros, mas podem, por outro lado, introduzir respostas socialmente desejáveis.

A escolha de um projeto de modos mistos também depende dos motivos para combiná-los: minimizar o erro total da pesquisa, reduzir custos ou aumentar a velocidade da coleta de dados. Se a velocidade for a consideração principal, pode-se oferecer todos os modos ao mesmo tempo, no chamado design simultâneo (*concurrent design*). Se os custos forem a principal consideração, deve-se empregar um design consecutivo, começando com o modo mais barato (*web*), e somente introduzindo outros modos após um período de tempo.

Como Medway e Fulton (2012) afirmam em uma meta-análise, que ao oferecerem a escolha entre o questionário *web* ou por carta, recebiam taxas de resposta levemente menores de quando ofereciam apenas o questionário pelo correio (Schouten, 2023). Schwartz (2004) *apud* Biemer et al. (2018) sugere que isso se dá pelos respondentes julgarem os benefícios e fardos de cada modo, tornando ambos menos atraentes, comparado à quando são oferecidos apenas um modo. O uso do design simultâneo também pode trazer custos adicionais não antes previstos, no que a maioria das pessoas optarão pela resposta em papel. Biemer et al. (2018) demonstra que ao oferecer incentivos monetários pela escolha da opção *web*, é possível aumentar consideravelmente a frequência de respostas online enquanto ainda diminuindo os custos totais da pesquisa. Ainda é incerto se pesquisas de design simultâneo possuem implicações e resultados diferentes nos dados ao compará-las à pesquisas consecutivas.

2.1 MEIOS DE CONTATO

A maioria dos institutos de pesquisa não têm acesso a registros de números de telefone ou endereços de correio eletrônico. Isto significa que, mesmo que as pessoas da amostra sejam convidadas a participar de um questionário na Internet, o meio mais comum de informar do pedido é por carta. Há um número cada vez maior de pesquisas sobre a melhor

forma de designs digitais (Dillman, 2017). Os elementos desta estratégia de comunicação são como descrever a tarefa de fazer login na *web*, como lembrar o entrevistado e com que frequência, como transmitir confiabilidade e se e como oferecer incentivos. Dillman et al. (2014) desenvolveram diretrizes para aumentar taxas de resposta da *web*:

- Utilizar múltiplos modos de contato com objetivo de aumentar a probabilidade de que convites para participar do questionário sejam recebidos e atendidos.
- Realizar o contato através de modos diferentes, com objetivo de aumentar a confiança do entrevistado na legitimidade da pesquisa. Assim, mesmo possuindo acesso ao número de celular e e-mail, é boa prática o envio adicional de convites por carta.
- Tentar obter informações de contato para mais de um modo de pesquisa.
- Enviar um pequeno incentivo monetário em conjunto com o primeiro contato por correio para aumentar a confiança na pesquisa.

Biemer (2018) afirma que um incentivo monetário inicial pode causar um aumento maior nas taxas de resposta ao se comparar à promessa de um bônus monetário após o envio das respostas, demonstrado por Avdeyeva e Matland (2013), que ao realizarem uma amostragem populacional geral na Rússia, descobriram que um incentivo incondicional (ou antecipado) de 50 rublos produzia taxas de resposta mais elevadas do que um incentivo condicional (ou prometido) de 300 rublos. No entanto, a oferta de ambos os tipos de incentivos produziu a taxa de resposta mais elevada de qualquer uma das alternativas testadas (Biemer, 2018). Acredita-se que tais incentivos aumentam a confiança na pesquisa. No entanto, Schouten (2023) aponta como os incentivos diferenciais entre os primeiros e os últimos entrevistados podem ser uma questão ética.

A crise da COVID, que pressionou os institutos de estatística a repensar a sua estratégia de recolha de dados, também os convidou a repensar a estratégia de comunicação. Beck et al. (2022) apud Schouten et al. (2023) descrevem que os NSIs europeus efetuaram alterações nos seus canais de contato, por exemplo, incluindo um código QR na carta convite, ou através da utilização de e-mails e mensagens de texto. Estas fontes de contato são especialmente utilizadas para lembretes e informações adicionais. Em alguns casos, a legislação nacional foi adaptada para dar aos NSIs o direito de acessar números de telefone e endereços de email disponibilizados por administrações públicas (como a autoridade fiscal). Mas obter o número ou o endereço de e-mail do entrevistado, seja solicitando-o na carta-convite, ou visitando o domicílio também foi frequentemente utilizado.

2.2 VIÉS DOS MODOS

Ao escolher o modo utilizado em uma amostra, seja ela de modo único ou modos mistos, é de suma importância a análise dos vieses provenientes do mesmo. Existem, na literatura, numerosos estudos com foco em comparar as estimativas obtidas vindas de duas estratégias amostrais diferentes, como comparar entrevistas por telefone com entrevistas cara a cara, ou questionários realizados online com aqueles feitos por carta. Um objetivo importante nesses estudos é determinar se as estimativas produzidas a partir dos diferentes modos são iguais e, caso contrário, determinar qual modo é melhor no sentido de fornecer vieses totais menores para os itens de interesse. Esses chamados estudos de comparação de modos geralmente envolvem um experimento de amostra dividida, onde cada subamostra aleatória é atribuída a um dos modos. Os dados são recolhidos utilizando essencialmente o mesmo questionário e procedimentos de entrevista. As estimativas são comparadas e quaisquer diferenças significativas entre as estimativas do mesmo parâmetro populacional são atribuídas a vieses decorrentes de um ou ambos os modos de entrevista.

Para alguns estudos de comparação, pode haver muito pouca informação disponível sobre as magnitudes dos vieses dos modos. Contudo, se houver informação disponível sobre a direção das tendências para ambos os modos, o modo menos tendencioso pode, por vezes, ser identificado. Por exemplo, para perguntas sobre temas delicados, como consumo de álcool e uso de drogas, às vezes supõe-se que o modo de entrevista que proporciona maiores valores nas variáveis referentes aos comportamentos sensíveis será menos tendencioso, uma vez que a tendência na população seria subestimar o comportamento ao relatar comportamentos sensíveis.

Nem sempre é possível utilizar o “senso comum” para escolher o modo menos viesado, o que exige que os pesquisadores incorporem características no desenho do estudo que produzam estimativas diretas do viés associado a cada modo. Isso costuma ser feito comparando os dados da amostra à dados externos que podemos assumir ser mais precisos do que as estimativas originais, tornando-os o “padrão-ouro” para o modo. Por exemplo, na Dinamarca, Kõrmendi (1988) *apud* Biemer (2001) estimou os vieses dos modos por telefone e entrevistas cara a cara utilizando dados da renda vindos das autoridades fiscais.

Biemer (2001) entretanto cita diversas limitações ao uso de dados externos como registros administrativos para avaliar a precisão dos modos utilizados. Algumas destas são: a indisponibilidade de dados externos para todas as variáveis de interesse; diferenças nas definições da variável na pesquisa e nas disponíveis nos registros; diferenças nos períodos de

tempo cobertos pela pesquisa e pelos dados externos; e ambiguidades na correspondência dos entrevistados com os registros administrativos. Essas limitações podem levar a vieses nos efeitos do modo estimado, o que pode resultar em conclusões enganosas sobre qual é o modo mais preciso.

Além disso, uma vez que o viés total atribuível a um modo de entrevista é, na verdade, a soma dos vieses de uma série de fontes de erro, interpretar os resultados das comparações de vieses totais é arriscado e potencialmente enganoso. Por exemplo, dois importantes componentes do viés total são o viés de medição e o viés de não resposta. Como visto por Biemer (2001), esses vieses podem ser compensadores; isto é, o sinal do viés de não resposta pode ser oposto ao sinal do viés de medição, resultando em um viés geral pequeno das duas fontes combinadas. Assim, um pequeno viés total estimado poderia ocultar vieses substanciais dos componentes.

2.3 DESIGN DO QUESTIONÁRIOS

A administração de questionários em modos diferentes corre o risco de introduzir diferenças de medição aos vieses já existentes. A maneira como os entrevistados entendem e respondem às perguntas pode ser afetada por características como a presença ou ausência de um entrevistador, se as perguntas são ouvidas ou lidas e a forma como as perguntas são apresentadas visualmente em uma tela. O design do questionário de modos mistos é uma ferramenta importante para prevenir ou reduzir os efeitos de medição específicos do modo.

Idealmente, as propriedades do questionário desempenhariam um papel importante na concepção da pesquisa. Por exemplo, se for identificado na fase conceitual que serão feitas principalmente perguntas sensíveis, é aconselhável não misturar modos de entrevista com modos sem entrevistadores. Uma alternativa, caso a presença do entrevistador seja essencial, é a utilização de ferramentas de auto-entrevista, nas quais o entrevistador não teria acesso às respostas dos entrevistados, mas ainda poderia oferecer apoio caso necessário (Schouten et al. 2023).

Uma preocupação que limita o interesse na amostragem baseada em endereços com questionários em papel e/ou na Internet foi o fato de as respostas encontradas nos questionários *web* serem provavelmente diferentes das respostas por telefone e face a face. Dois aspectos ditam essa preocupação. O primeiro é que, sem um entrevistador, os entrevistados não podem receber incentivo extra quando forem incapazes ou relutantes em

responder a uma pergunta, nem pode ser corrigido algum mal-entendido nas perguntas. O segundo é a evidência de longa data de que a deseabilidade social e a tendência por respostas mais socialmente aceitas são maiores para respostas de pesquisas por telefone e cara a cara do que para respostas autoadministradas (como correio ou *web*) (de Leeuw 1992 apud Dillman 2017). Tradicionalmente, o benefício de ter um entrevistador presente é visto como algo que supera o potencial viés do segundo fator em amostras de modo único.

Questionários aplicados por modos mistos podem ser elaborados com foco em minimizar as diferenças de medição entre os modos, apresentando um estímulo unificado em todos os modos utilizados. Isto é chamado de abordagem de modo unificado. Alternativamente, a “abordagem de melhores práticas” centra-se na minimização de erros dentro de cada modo, mesmo que isso signifique usar estímulos diferentes em modos diferentes.

Na prática, as comparações entre grupos e ao longo do tempo são fundamentais para a maioria das pesquisas. Portanto, o design do questionário de modos mistos geralmente se concentra em minimizar as diferenças de medição entre os modos, ou seja, a abordagem de modo unificado. Também do ponto de vista prático e de eficiência de custos é preferível conceber questionários para vários modos tão idênticos quanto possível. No entanto, é notável que algumas diferenças entre os modos devem necessariamente existir, mesmo em uma abordagem unificada (por exemplo, “pressione Enter para continuar”).

A utilização do modo unificado pode ser reforçada ao analisar vieses característicos dos designs de cada modo. Christian, Dillman e Smyth (2008), como apresentado por Schouten (2023), demonstram forte evidência de que respostas por telefone tendem a produzir respostas em pólos extremos ao tratar de perguntas utilizando quantitativos vagos, e menos uso de categorias intermediárias, ao serem comparadas às respostas por cartas ou *web*. A razão aparente para esta diferença é que a apresentação visual das categorias de respostas intermediárias em questionários *web* ou em papel é mais perceptível e, portanto, acessível aos entrevistados do que quando essas mesmas categorias são lidas por telefone, um processo que torna as categorias finais mais proeminentes nas mentes dos entrevistados (Dillman e Edwards 2016 *apud* Schouten et al. 2023).

Schouten et al. (2023) não aconselham a adição de um novo modo, como questionários *web*, a uma pesquisa existente. No lugar, o correto é realizar uma reformulação total do questionário se um novo modo for adicionado. Eles também aconselham que os requisitos de modos mistos devem ser considerados em cada etapa do processo de elaboração do questionário, por exemplo, o pré-teste do questionário e a avaliação do trabalho de campo.

3. ANÁLISE COMPARATIVA DE QUESTIONÁRIOS POR WEB E TELEFONE

É importante, para o objetivo deste trabalho, entendermos os benefícios e desafios do uso de questionários via *web* e por telefone.

A utilização de métodos de coleta de dados via *web* tem sido incentivada por uma série de considerações, sobretudo os problemas aparentemente insolúveis de pesquisas por telefone e correio. Fundamentalmente, contatar domicílios ou indivíduos através de um modo, como correio ou telefone, para solicitar que respondam por outro modo, não é um procedimento ideal de coleta de dados. É provável que haja algum atrito entre receber uma carta postal ou um telefonema e depois ter que mudar para um modo de resposta diferente. A mudança por si só provavelmente afetará as taxas de resposta. Assim, não é surpreendente que as dificuldades na realização de inquéritos monomodo por telefone e por e-mail/*web* sejam a razão fundamental para procurar uma alternativa.

3.1 A DECRESCENTE EFICÁCIA DE PESQUISAS TELEFÔNICAS

Embora pesquisas telefônicas tenham sido utilizadas ocasionalmente para apoiar a coleta de dados, o desenvolvimento do telefone como único meio de coleta de respostas a pesquisas só ocorreu no início da década de 1970 (Nathan 2001 *apud* Dillman 2017). A utilização de métodos de coleta de dados telefônicos avançou rapidamente devido à crescente presença do telefone nos domicílios familiares e ao desenvolvimento do procedimento Mitofsky-Waxberg para a utilização de métodos de discagem aleatória de dígitos na seleção de lares. Além disso, a diminuição dos custos das chamadas de longa distância resultou em pesquisas por RDD (*random digit dialing*) por telefone substituírem a maior parte das entrevistas presenciais (Dillman 2017).

Durante os anos, numerosas pesquisas foram realizadas para avaliar a eficácia de amostragens por telefone utilizando o método RDD. Entre 1997 e 2012, como demonstrado por Dillman (2017), o Pew Research Center (2012), um importante realizador de inquéritos sociais por telefone nos Estados Unidos, reportou quedas nas taxas de resposta por RDD de 35 por cento para cerca de 9 por cento. Mais recentemente, Dutwin e Lavrakas (2016) *apud* Dillman (2017) conduziram uma análise das taxas de resposta telefônica para nove organizações. Eles descobriram que as taxas de resposta de telefones fixos diminuíram de 15,7 por cento em 2008 para 9,3 por cento em 2015, enquanto as taxas de resposta de

telefones celulares diminuíram durante este período de 11,6 para 7,0 por cento. Eles também relataram que esse declínio de cerca de 40% na resposta é menos resultado de um aumento nas recusas do que de um aumento de não respostas e secretárias eletrônicas de 10 pontos percentuais para telefones fixos e 24 pontos percentuais para telefones celulares.

O telefone deixou de ser um dispositivo doméstico, compartilhado por todos os membros de um mesmo domicílio, para se tornar um instrumento sem fios de posse individual, facilmente transportável de um local para outro. No Brasil, apenas 9,5% de todos os domicílios familiares ainda utilizam redes fixas, enquanto 96,7% apresentam redes sem fio. (IBGE, 2023). Uma implicação do aumento da proporção de aparelhos celulares é que a amostragem domiciliar se tornou muito mais difícil. É possível incluir números de celular em quadros de amostra RDD, contudo, tornou-se também necessário dedicar preciosos minutos de entrevista para averiguar uma série de informações, incluindo o número e o tipo de telefone numa casa, a fim de determinar as probabilidades de seleção do domicílio familiar.

Os telefones fixos e celulares enfrentam conjuntamente um desafio maior. Cada vez menos pessoas participam de conversas de voz por telefone. Esta é uma grande mudança desde a época em que a comunicação essencial para discussões de negócios, a manutenção de relações sociais e a coordenação de atividades diárias eram feitas principalmente por chamadas de voz. E-mail e mensagens de texto substituíram majoritariamente esse uso. Falar ao telefone com um entrevistador de pesquisa está cada vez mais fora da realidade comparado a outros aspectos da vida diária das pessoas.

O uso contínuo do telefone como ferramenta para coleta de dados é incentivado pelo grande investimento que as organizações têm em hardware, software e pessoal especializado, muitos dos quais não haviam realizado outros tipos de coleta de dados. No entanto, o declínio contínuo das taxas de resposta telefônica nos últimos anos, observado por Dutwin e Lavrakas (2016) *apud* Dillman (2017), e as preocupações com o valor dos resultados obtidos reduziram a credibilidade da realização de pesquisas telefônicas como modo único.

3.2 VANTAGENS ESTRATÉGICAS DE PESQUISAS POR TELEFONE

Enquanto pesquisas por telefone podem não ser as mais atrativas ao realizar estudos diretamente com o público, é notável sua capacidade de atingir pequenos grupos populacionais nos quais outros modos de coleta possuem dificuldade de alcançar (THE FARNSWORTH GROUP, 2024).

A acessibilidade dos respondentes ao questionário também se torna um ponto importante para a amostra, principalmente ao lidarmos com uma população que não seja tão

proficiente com tecnologia ou possua acesso à internet. Seria quase impossível realizar uma pesquisa online com objetivo de coletar dados de habitantes de zonas rurais, enquanto o acesso ao telefone nestas áreas já é mais expandido (CETIC, 2023).

Este método de pesquisa não só permite encontrar e conversar com uma parcela da população que não é muito ativa em canais online, como também garantir melhor a qualificação de cada respondente. Como o entrevistador está contatando indivíduos diretamente para obter sua participação, ele pode garantir que eles atendam aos seus critérios, diminuirmos consideravelmente a chance de respostas realizadas por terceiros em nome do entrevistado pretendido.

Pesquisas telefônicas também permitem uma melhor administração e feedback de perguntas qualitativas que são difíceis de captar através de métodos de pesquisa com a ausência de um entrevistador. De acordo com Jason Anderson, gerente sênior de trabalho de campo do The Farnsworth Group (2024), “Para perguntas abertas e coleta de dados qualitativas, você pode ir além da primeira resposta do entrevistado e fazer perguntas de acompanhamento para obter uma melhor compreensão de seu respostas e uma melhor elaboração de suas ideias”. Se uma pergunta não for formulada com clareza ou se os entrevistados não a entenderem completamente, o entrevistador receberá feedback em tempo real sobre essas questões, bem como uma oportunidade de fornecer mais explicações ou clareza para coletar com sucesso as informações desejadas.

O fato dos entrevistadores estarem em tempo real coletando as respostas dos entrevistados, eles garantem também que tais indivíduos não percam o interesse ou se cansem durante a aplicação do questionário, um problema recorrente em pesquisas online. Em geral, conversar diretamente com o indivíduo que está sendo entrevistado é mais envolvente. O entrevistador pode controlar o processo de coleta de dados e garantir que cada entrevistado permaneça atento e dedique algum tempo para responder às perguntas. Isto gera uma coleta de dados de maior qualidade (THE FARNSWORTH GROUP, 2024).

A velocidade com que os dados podem ser recolhidos e processados através de pesquisas telefônicas é muito mais rápida do que quando comparada com entrevistas presenciais ou online, em que o pesquisador deve aguardar um período determinado de tempo para que os respondentes lhe enviem suas respostas. O surgimento do sistema CATI (*Computer-assisted Telephone Interviewing*) também ajuda a acelerar a coleta, o processamento e o gerenciamento de dados, tornando o processamento, organização e armazenamento dos dados da pesquisa muito mais fácil para os entrevistadores. Ele

automatiza tarefas como discar um número, salvar dados e preencher detalhes adicionais, o que economiza tempo do entrevistador e agiliza a coleta de dados.

3.3 SMARTPHONES E A FALTA DE CONFIANÇA DO PÚBLICO, DESAFIOS DE PESQUISAS ONLINE

Em meados da década de 1990, quando as taxas de resposta de pesquisas por telefone iniciavam o seu declínio persistente, pesquisas pela internet, o substituto esperado, iniciavam o seu rápido desenvolvimento (Dillman 2017). No entanto, duas décadas mais tarde, a sua utilização para amostras populacionais em geral continua limitada.

Os endereços de e-mail não existem em formatos padrão como é o caso dos nossos números de telefone que identificam um código de área, uma central e as limitadas possibilidades de números em cada central. É também provável que as pessoas dentro dos agregados familiares tenham vários endereços de e-mail, de modo que as probabilidades de alcançar agregados familiares específicos ou outras unidades amostrais não podem ser calculadas. Além disso, parte da população com maior conhecimento de informática, adolescentes e jovens adultos, desenvolveu uma reputação de minimizar o uso de sistemas de e-mail tradicionais. Em vez disso, eles se concentram fortemente em redes sociais como Instagram, no WhatsApp e em outros aplicativos de mensagens instantâneas para se conectar com amigos e conhecidos.

Além disso, as taxas de resposta de pesquisas *web* para amostras aleatórias de endereços de e-mail existentes são provavelmente tão baixas ou inferiores às alcançadas em pesquisas telefônicas atuais (Lozar et al. 2008 apud Dillman 2017), e é provável que incluam um número desproporcionalmente elevado de indivíduos mais jovens e com maior nível de escolaridade, apesar de muitos jovens dependerem de outras formas de ligação eletrônica que os tornam apenas utilizadores ocasionais do correio eletrônico tradicional. As caixas de entrada dos computadores das pessoas são normalmente um espaço lotado, com e-mails não solicitados e indesejados sendo mais prevalentes do que o número de chamadas telefônicas indesejadas. Além disso, os e-mails são frequentemente verificados e excluídos com base apenas na origem ou após a leitura de apenas algumas palavras iniciais da mensagem que os acompanha (Dillman 2017). Isso ainda sem considerarmos os sistemas automáticos das caixas de entrada que muitas vezes excluem ou direcionam e-mails de fontes desconhecidas para caixas de Spam, fazendo com que não sejam sequer analisados brevemente.

As mudanças nas tecnologias da informação e comunicação também contribuem para a falta de resposta às pesquisas na *web*. Na era dos smartphones, quando as pessoas têm

maior probabilidade de se deslocar de um lugar para outro, a concentração no conteúdo das pesquisas parece ser menos provável de ser alcançada. É evidente que a proporção de questionários concluídos através de smartphones está a aumentar em relação a todas as pesquisas online (Couper, Antoun e Mavletova 2017 *apud* Dillman 2001). No entanto, parece não haver evidências de que questionários respondidos através de smartphones estejam a aumentar as taxas de resposta totais das pesquisas, podendo de facto estarem a reduzi-las. Os smartphones, por sua natureza, são dispositivos de interrupção, com possibilidade de receber textos, chamadas de voz e e-mails a qualquer momento, muitas vezes enquanto os indivíduos se movimentam fisicamente entre suas atividades diárias.

Com o avanço de ferramentas online, também é notável o crescimento nas tentativas de pessoas mal intencionadas de usarem essas mesmas ferramentas para uso próprio. Notícias de ataques de hackers e invasões em bancos de dados de grandes empresas são recorrentes em nosso dia a dia. O resultado dessas diversas possibilidades é a preocupação de muitas pessoas em relação à sua segurança, ou a falta dela, do site e com as informações que fornecem em resposta a solicitações de pesquisas na *web*. A falta de confiança em questionários online e a preocupação de que a informação possa ser guardada e utilizada para fins não relacionados com a pesquisa são também potenciais barreiras à resposta. É difícil prever se os avanços tecnológicos e de controle social irão anular os riscos associados à utilização de computadores. Por enquanto, esta é uma questão que ameaça o sucesso de pesquisas pela internet e que não pode ser ignorada.

3.4 MOTIVOS PARA O OTIMISMO

O desenvolvimento e a implantação de metodologias *web* para a coleta de dados durante a última década fornecem razões para otimismo quanto à possibilidade de amostras de maior qualidade. Esse otimismo surge menos do entusiasmo com uma abordagem específica de contactar as pessoas e convencê-las a responder na *web* do que de uma combinação de considerações. O uso de pesquisas *web* não vem de um desejo de utilizá-las singularmente em amostragens de grande escala, mas como uma ferramenta a ser adicionada ao arsenal de pesquisadores e pesquisas já existentes.

A utilização de pesquisas online vem naturalmente interligada com modos mistos de pesquisa. Como demonstrado por Dillman (2017), diversas agências de pesquisa ao redor do mundo já utilizaram ou ainda utilizam questionários online. Por exemplo, o American Community Survey, que serve como a principal fonte de informação estatal e regional sobre os agregados familiares dos EUA, começou a utilizar uma abordagem *web* para a coleta de

dados em 2013, que incluiu também a possibilidade de responder mais tarde no processo de implementação por correio, telefone ou entrevista presencial. Outros exemplos incluem o Censo Japonês de 2015 e os Censos de 2016 no Canadá e na Austrália. Além disso, a Pesquisa de Graduados Universitários dos EUA, realizada a cada 2-3 anos pela National Science Foundation, completou a mudança da coleta de dados por correio e telefone para uma abordagem digital, seguida pelos outros dois modos de coleta de dados (Finamore e Dillman 2013 *apud* Dillman 2017). Estes exemplos são apenas alguns dos principais esforços de pesquisa em todo o mundo que utilizam agora esta metodologia.

De acordo com a Casa Civil (2022), 90% dos lares brasileiros já têm acesso à internet, o que é mais elevado do que foi para telefones quando o rápido desenvolvimento das pesquisas por telefone ocorreu no início da década de 1970 (Nathan 2001 *apud* Dillman 2017). As competências de utilização da Internet são hoje em dia fundamentais para o processo educativo, para as operações organizacionais e para o acesso aos serviços de consumo. A crise sanitária do COVID-19 mostrou que dadas as ferramentas e circunstâncias necessárias, é possível estender o alcance da *web* até àqueles menos favorecidos ou em zonas de difícil acesso. Da mesma forma que jovens e adultos foram possibilitados de participar de aulas e cursos pela *web*, as pesquisas online podem ser utilizadas para alcançar um público que normalmente não seria facilmente incluído em pesquisas presenciais ou por telefone.

As descobertas sobre como a informação visual é processada, relatadas por Palmer (1999), Hoffman (2004) e Ware (2004), como apontadas por Dillman (2017), forneceram visões teóricas sobre as ações separadas que ocorrem à medida que o olho capta a informação e o cérebro a processa para dar sentido ao que está acontecendo na página ou tela. A aplicação destes conceitos proporcionou uma compreensão das razões pelas quais os questionários auto administrados frequentemente produzem respostas diferentes das pesquisas por entrevista. Entrevistados são normalmente guiados através de questionários visuais por múltiplas linguagens que comunicam significado. Eles incluem símbolos, números e sua composição gráfica (tamanho, espaçamento, cor, simetria, regularidade, etc.) que afetam como as informações em papel e páginas da *web* são navegadas, agrupadas mentalmente e interpretadas (Tourangeau, Couper e Conrado 2004 *apud* Dillman 2017). Pesquisas adicionais mostraram que a conformidade com as instruções de ramificação poderia ser melhorada drasticamente por meio de mudanças nos símbolos, no tamanho da fonte, no brilho da fonte (Redline e Dillman 2002; Christian e Dillman 2004 *apud* Dillman 2017) e no posicionamento dessas instruções de ramificação em relação às escolhas de resposta. Isso demonstra uma natureza muito mais moldável e com inúmeros potenciais de

crescimento para futuras pesquisas *web*, algo de muito mais difícil confecção quando tratamos de pesquisas por correio, telefone ou pessoais. Além disso, a utilização em grande medida da autoadministração (Internet e correio) proporciona uma melhor adaptação cultural às pessoas do que uma conversa telefônica, que está cada vez mais fora de sincronia com o comportamento de comunicação rotineiro que dá grande ênfase às mensagens de texto e ao e-mail.

Uma das vantagens mais significativas das pesquisas por email ou ferramentas online é a redução de custos. Diferentemente das entrevistas telefônicas, que podem incorrer em altos custos de chamadas e necessidade de diversos operadores, ou pesquisas por cartas que exigem milhares de folhas de papel, as pesquisas por email eliminam a necessidade de ligação direta e reduzem os custos operacionais associados. Além disso, não há necessidade de impressão ou envio físico de questionários, o que também reduz despesas consideráveis. Com a pesquisa online existem ferramentas muitas vezes gratuitas ou de baixo custo para conseguir enviar formulários. Ganhando desta forma um melhor custo benefício e também sustentabilidade, já que não serão gastos mais várias folhas de papel para imprimir as pesquisas (Mininel, 2021). Este custo-benefício se estende também aos subsequentes contatos que podem ser necessários ao longo da pesquisa, que tornarão a pesquisa exponencialmente mais barata de acordo com o quão grande ou duradoura ela seria.

4. REFERENCIAL TEÓRICO

Seguindo as proposições de Cochran (1977), temos a ideia de que a utilização de amostragem estratificada é uma maneira efetiva de aumentar a precisão de estimativas, assim como permite trabalhar com tamanhos de amostra inferiores para um número de precisão fixo. Adicionalmente, Montgomery e Runger (2012) abordam a importância e a aplicação dos intervalos de confiança na comparação de métodos de coleta de dados, oferecendo uma compreensão aprofundada sobre como esses intervalos ajudam a garantir a precisão e a representatividade das estimativas. Ambos métodos são utilizados em conjunto na pesquisa com objetivo de alcançar maior precisão na comparação entre as pesquisas por e-mail e por telefone.

4.1 AMOSTRAGEM ESTRATIFICADA

A Amostragem Estratificada (AE) envolve a divisão da população em estratos mutuamente exclusivos e exaustivos, de acordo com características conhecidas. São

selecionadas unidades de forma independente de cada estrato, com subgrupos mais homogêneos que a população como um todo, com objetivo de proporcionar uma redução do erro amostral (Bolfarine e Bussab, 2005).

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

De acordo com Kish (1995), a média populacional (\bar{Y}_h) dentro de cada estrato pode ser calculada ao dividir o total dentro daquele estrato (Y_h) pelo tamanho do estrato (N_h). Adicionalmente, a média populacional é dada por

$$\bar{Y} = \sum_{h=1}^H W_h \bar{Y}_h \quad \text{onde} \quad W_h = N_h/N$$

A partir disso, sabe-se também que a média amostral ponderada é dada por

$$\bar{y} = \sum_{h=1}^H W_h \bar{y}_h$$

A média amostral estimada dentro do estrato é dada por

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

A variância dessa média é

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

O estimador da variância da média amostral ponderada é dado por

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

O estimador do erro padrão da média amostral ($ep(\bar{y})$) é encontrado ao calcular a raiz quadrada positiva do estimador da variância $\widehat{V}(\bar{y})$ (Montgomery 2012).

Para o cálculo da proporção populacional e amostral dentro de cada estrato, entretanto, devemos calcular $P_h = \frac{X_h}{N_h}$, e $\hat{p}_h = \frac{x_h}{n_h}$, onde X_h e x_h são o número de sucessos no estrato h na população e na amostra, respectivamente. Logo, as proporções da população e da amostra estratificadas são dadas por uma média ponderada das proporções dentro dos estratos:

$$P = \sum_{h=1}^H W_h P_h \quad \text{e} \quad \hat{p} = \sum_{h=1}^H W_h \hat{p}_h$$

O estimador da variância da variância amostral estratificada é dado por

$$\widehat{V}(\hat{p}) = \sum_{h=1}^H W_h^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h}$$

De forma semelhante ao estimador do erro padrão da média amostral, o estimador do erro padrão da proporção amostral ($ep(\hat{p})$) é encontrado ao calcular a raiz quadrada positiva do estimador da variância $\widehat{V}(\hat{p})$ (Montgomery 2012).

4.2 INTERVALOS DE CONFIANÇA

Um intervalo de confiança é uma faixa de valores calculada a partir de dados amostrais que se destina a estimar um parâmetro populacional desconhecido. Este intervalo é construído de tal forma que, com um certo nível de confiança (comumente 95% ou 99%), espera-se que este intervalo contenha o parâmetro populacional verdadeiro (Montgomery, 2012). Em outras palavras, um intervalo de confiança fornece uma estimativa, ou aproximação, da média amostral. Diferentemente de um simples ponto estimado, o intervalo de confiança oferece uma faixa de valores plausíveis, proporcionando uma visão mais completa e precisa da incerteza associada à estimativa amostral. Além disso, ao considerarem a variabilidade dos dados e o tamanho da amostra, os intervalos de confiança ajudam a avaliar a confiabilidade dos resultados e a robustez das conclusões tiradas. Em um caso onde

a média populacional é conhecida, o intervalo de confiança ainda pode ser utilizado para averiguar a precisão de uma ou mais amostras (Montgomery, 2012).

Para o cálculo do intervalo de confiança para a média populacional, calcula-se a média amostral (\bar{y}) e o estimador do erro padrão desta mesma média amostral ($ep(\bar{y})$). Em seguida, deve-se selecionar o nível de confiança desejado, como 95% ou 99%, e encontrar o valor crítico correspondente da distribuição t de Student (para amostras pequenas e variância desconhecida) ou da distribuição normal (para amostras grandes). Na prática, uma amostra é considerada grande quando seu tamanho é maior que 30 (Triola, 2013). O intervalo de confiança é então dado por

$$\bar{y} \pm (z * ep(\bar{y})) \quad \text{ou} \quad \bar{y} \pm (t * ep(\bar{y}))$$

Onde z e t são os valores críticos das distribuições Normal e t de Student, respectivamente. Este cálculo resulta em uma faixa de valores dentro da qual espera-se que a média populacional verdadeira esteja contida, proporcionando uma medida de precisão da estimativa amostral (Triola, 2013; Ross, 2013).

De forma semelhante, ao calcular o intervalo de confiança para uma proporção populacional, calcula-se a proporção amostral (\hat{p}) e o estimador do erro padrão desta mesma proporção amostral ($ep(\hat{p})$). Para o nível de confiança e a distribuição escolhida, podem ser seguidos os mesmos parâmetros impostos para o intervalo de confiança da média amostral (Scheaffer 2011). O intervalo de confiança é então dado por

$$\hat{p} \pm (z * ep(\hat{p})) \quad \text{ou} \quad \hat{p} \pm (t * ep(\hat{p}))$$

5. MÉTODOS

Antes de prosseguir com o estudo, foi essencial compreender o funcionamento do Censo Escolar do INEP.

5.1 CENSO ESCOLAR DO INEP

O Censo Escolar, conduzido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), é a principal fonte de informações sobre a educação básica no Brasil. Iniciado em 1990, o Censo Escolar tem suas raízes no compromisso do

governo brasileiro com a melhoria da qualidade da educação e a garantia de acesso à educação para todos. Este compromisso foi consolidado com a Constituição Federal de 1988, que estabeleceu a educação como um direito de todos e um dever do Estado (BRASIL, 1988).

O Censo Escolar é realizado anualmente e abrange todas as escolas de educação básica do país, tanto públicas quanto privadas. Ele coleta dados sobre diversos aspectos da educação, incluindo matrícula, infraestrutura escolar, formação de professores e desempenho dos alunos. A coleta de dados é feita em duas etapas: Na coleta inicial, os dados são coletados diretamente das escolas por meio de um sistema online denominado Educacenso. Os gestores escolares são responsáveis por inserir as informações no sistema, que incluem detalhes sobre alunos, turmas, docentes e recursos escolares (INEP, 2022). Após a coleta inicial, os dados passam por um processo de validação e verificação realizado pelo INEP em parceria com as secretarias estaduais e municipais de educação. Este processo visa assegurar a precisão e a integridade dos dados coletados (INEP, 2022).

Os dados do Censo Escolar são fundamentais para o planejamento e a implementação de políticas educacionais no Brasil. Eles são utilizados pelo Ministério da Educação (MEC) e outras entidades governamentais para a formulação de estratégias de melhoria da qualidade da educação e para a distribuição de recursos financeiros através de programas como o Fundo de Manutenção e Desenvolvimento da Educação Básica (FUNDEB) (MEC, 2023). Além disso, os dados são acessíveis ao público, permitindo que pesquisadores, gestores e a sociedade em geral acompanhem o panorama da educação no país e identifiquem áreas que necessitam de intervenção (INEP, 2022).

5.2 VARIÁVEIS

Para a realização deste estudo, escolhemos utilizar variáveis de livre acesso no Censo Escolar do INEP. Seguindo as ideias de Biemer (2001), a escolha do Censo Escolar como fonte de dados deve-se ao fato de ele ser considerado um padrão ouro para a análise da educação básica no Brasil, proporcionando dados precisos e amplamente reconhecidos (INEP, 2022). A utilização destas variáveis nos permite analisar a precisão das amostras que coletamos, garantindo que os dados obtidos sejam comparáveis e confiáveis. Um cuidado adicional, entretanto, teve de ser feito em relação à defasagem temporal que pode ocorrer devido à natureza anual da coleta de dados do Censo Escolar. Devido à periodicidade da coleta de dados, os dados podem não refletir as mudanças mais recentes nas condições

educacionais de escolas e regiões, especialmente em cenários de rápida mudança. Esse descompasso entre o momento da coleta do Censo Escolar e a coleta dos dados para a amostra pode impactar a precisão e a validade dos resultados. Para mitigar o problema, a coleta dos dados das amostras foi realizada próxima da data de postagem oficial dos dados do censo.

A seleção das variáveis foi feita com base em dois critérios principais: a facilidade de resposta por parte das escolas e a simplicidade de análise. Buscamos variáveis que todas as escolas conseguiram responder de maneira consistente e que fossem relevantes para a nossa pesquisa. Como nossa pesquisa tem como objetivo apenas comparar os modos de coleta de dados e amostragem, não foi necessário a procura de variáveis que seguissem um contexto específico ou que demonstrassem a qualidade da escola em si. As variáveis escolhidas foram: Números de Docentes de Educação Infantil (QT_DOC_INF), Números de Docentes de Ensino Fundamental (QT_DOC_FUND), Números de Docentes de Ensino Médio (QT_DOC_MED), Números de Turmas de Educação Infantil (QT_TUR_INF), Números de Turmas de Ensino Fundamental (QT_TUR_FUND), Números de Turmas de Ensino Médio (QT_TUR_MED), Existência de Biblioteca na Escola (IN_BIBLIOTECA), Existência de Laboratório de Informática na Escola (IN_LABORATORIO_INFORMATICA), Existência de Internet na Escola (IN_INTERNET), Existência de Cozinha na Escola (IN_COZINHA), Existência de Piscina na Escola (IN_PISCINA), Existência de Auditório na Escola (IN_AUDITORIO).

5.3 COLETA E ANÁLISE DOS DADOS

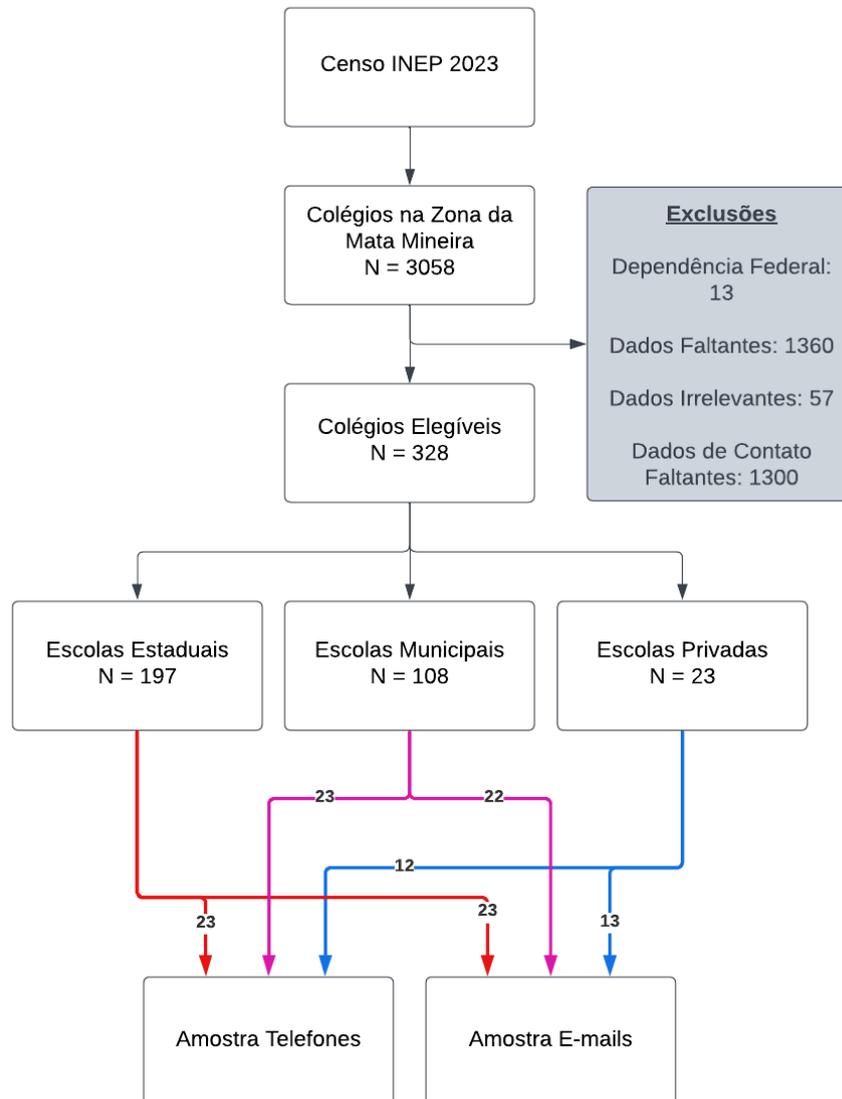
Antes de prosseguir para a coleta dos dados, foi necessário a utilização de uma segunda fonte que nos disponibilizasse os e-mails das escolas desejadas. O site utilizado foi o da Secretaria de Educação de Minas Gerais, que disponibiliza um cadastro de uma grande parcela das escolas de Minas Gerais, que então pode ser mesclado com os dados que já possuímos do Censo Escolar do INEP para adquirirmos nossa base de dados final. Sabendo que todas as escolas possuem um código específico dado pelo próprio governo, e que tanto o Censo Escolar como o Cadastro de Escolas da Secretaria de Educação utilizam disso para categorizar as escolas, podemos utilizar essa variável para apontarmos cada e-mail à sua respectiva escola no Censo Escolar. Como o Censo Escolar já nos disponibiliza o número de telefone e o DDD das escolas, temos todos os dados para prosseguir para a modelagem de nossa população. Para a importação, leitura e salvamento dos microdados do Censo Escolar

foi usado o pacote *readODS*, na interface R Studio, no software livre R, desenvolvido por Chan e Schutten (2023). Suas funções permitem baixar dados diretamente em formato de planilhas salvas em seu computador para uma melhor e mais fácil análise e manipulação dos dados.

Após a coleta dos dados, foi necessária uma limpeza para retirarmos todas as escolas não desejadas de nosso banco de dados, como demonstrada na figura 1. Inicialmente selecionamos apenas as escolas com a categorização Zona da Mata na variável `NO_MESORREGIÃO`, o que nos permitiu separar apenas as escolas da região pretendida para o estudo. Após isso, foram realizadas limpezas sequenciais para retirar dados desnecessários para nossa análise. Primeiramente, foram retiradas do banco de dados as escolas de dependência federal, devido a baixa quantidade de escolas no estrato, as quais apenas causariam análises de qualidade inferior. Posteriormente foram retiradas todas as escolas que possuíam dados faltantes, como as variáveis utilizadas nesse estudo ou seu número de telefone. Também foram retiradas todas as escolas que possuem apenas níveis educacionais irrelevantes para nossa pesquisa, estes sendo educação especial, educação profissional e educação de jovens e adultos. Por fim, foram utilizados os códigos das escolas para mesclar o banco de dados com os e-mails ao banco de dados do Censo Escolar, retirando qualquer escola com dados de contato faltantes. Isso fez com que o tamanho de nossa população final fosse reduzido consideravelmente para um total de 328 escolas.

Para o cálculo do tamanho da amostra foi calculado um tamanho de amostra aleatória simples sem reposição para estimação de uma proporção, considerando proporção esperada na população de 0,5, margem de erro de 7,5% e nível de confiança de 95%, como recomendado por Montgomery (2012). O tamanho amostral alcançado foi de 113 escolas. Em seguida, a amostra foi dividida igualmente, com o tamanho das duas amostras sendo arredondados para 57 escolas, onde uma das amostras foi usada para a coleta de dados por e-mail, enquanto a outra foi utilizada para a coleta por telefones.

Figura 1 – Fluxograma do estudo, com estratos e exclusões da população



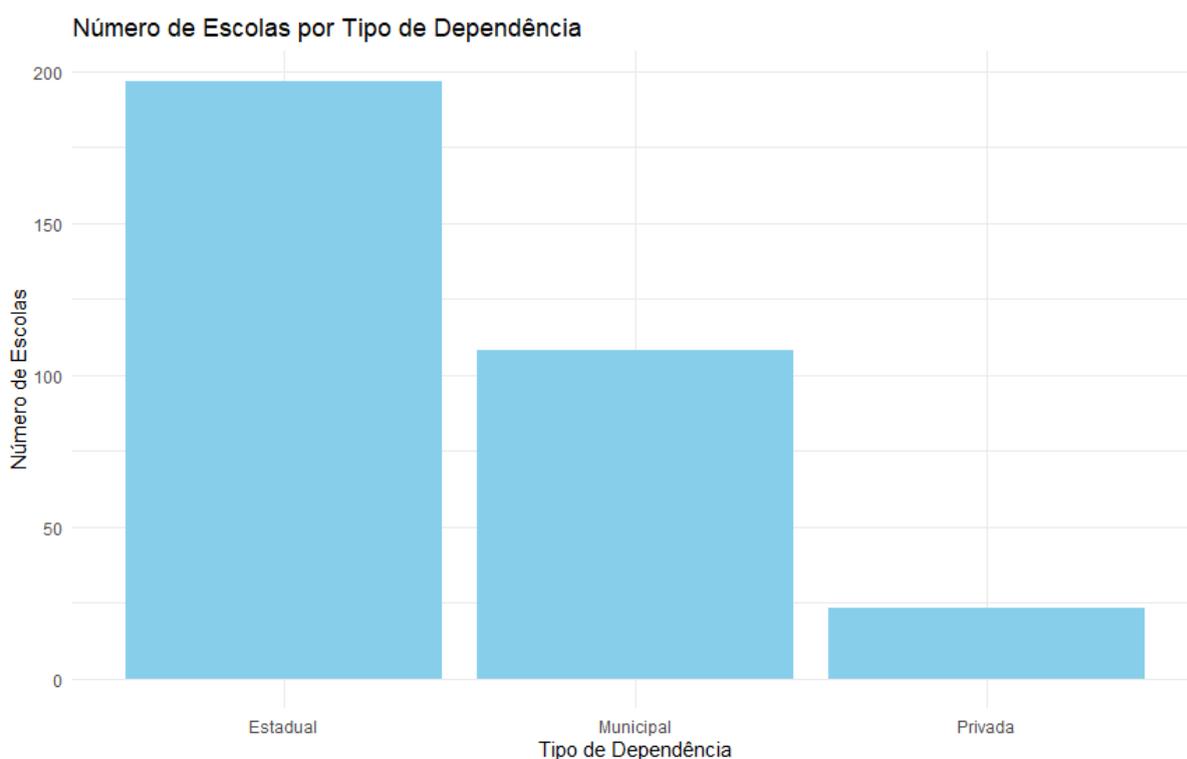
Fonte: do próprio autor (2024)

Para este estudo, foi utilizado o método de amostragem estratificada, com o objetivo de alcançar uma maior precisão das estimativas considerando o tamanho de nossa população. Este método foi escolhido também devido às restrições de tempo e recursos disponíveis, visto que a coleta de dados foi realizada de forma individual. A população foi dividida em 3 estratos, como visto na figura 1, com base na variável TP_DEPENDENCIA, na qual referenciava a dependência administrativa das escolas entre Estadual, Municipal e Privada.¹

¹ Originalmente a variável também levava em conta escolas de dependência federal, mas devido a baixa quantidade de escolas no estrato, foi considerado como melhor opção a retirada destas escolas da população.

Devido ao baixo número de escolas de dependência privada na população, não foi possível dividir os tamanhos dos estratos igualmente. No lugar, toda a população de escolas de dependência privada participou das amostras, sendo separadas treze escolas privadas para a coleta de dados por e-mail e doze para a coleta de dados por telefone. Para o restante das amostras, foram separados igualmente entre escolas estaduais e municipais, com vinte e três escolas estaduais e vinte e duas escolas municipais sendo sorteadas para a coleta de dados por e-mail, e vinte e três escolas estaduais e vinte e três escolas municipais sendo selecionadas para a coleta de dados por telefone. A proporção de escolas por tipo de dependência pode ser vista no gráfico 1.

Gráfico 1 – Número de Escolas da População por Tipo de Dependência



Fonte: do próprio autor (2024)

Para a coleta de dados da amostra de telefones, foram realizadas ligações para os números disponibilizados pelo Censo Escolar durante três semanas, em dias intercalados, com objetivo de entrar em contato com o diretor das escolas escolhidas, ou algum auxiliar com as informações desejadas. As respostas foram salvas manualmente pelo entrevistador utilizando a ferramenta de formulários do Google. Caso não houvesse alguém presente para responder o questionário, ou na situação em que o telefonema não era atendido, tentativas

subsequentes de contato foram realizadas em dias da semana diferentes e horários variados, com o objetivo de encontrar um respondente em um horário disponível. Após a primeira leva de telefonemas, foram adquiridos os dados completos de dezoito escolas da amostra, além de duas escolas nas quais desejaram responder apenas uma fração do questionário. Adicionalmente, um total de 13 escolas da amostra não foram alcançadas devido a falhas em seus números de contato, que não permitiram que ligação alguma fosse realizada. Foram realizadas então mais três tentativas de contato para o restante das vinte e quatro escolas da amostra, das quais apenas uma respondeu por completo o questionário após duas ligações. Das restantes, quatorze escolas não foram alcançadas, seis recusaram participar da pesquisa e três responderam o questionário de forma incompleta. Das quatorze escolas que não foram alcançadas, sete nunca atenderam o telefone, enquanto as outras sete protelaram o atendimento, constantemente adiando o horário de retorno e nunca estando disponíveis para completar a entrevista. Ao fim do período de coleta, o tamanho final da amostra (n) alcançado pela amostra de telefones foi de dezenove escolas. Entre as dezenove escolas, nove eram de dependência estadual, sete de dependência municipal e três de dependência privada. Para as variáveis categóricas como IN_BIBLIOTECA e IN_INTERNET, o tamanho final da amostra (n) alcançado pela amostra foi de vinte e três escolas, devido às escolas que responderam o questionário de forma incompleta. Entre as vinte e três escolas, onze eram de dependência estadual, sete de dependência municipal e cinco de dependência privada.

Para a coleta de dados da amostra de e-mails, foram enviados e-mails personalizados para cada um dos endereços disponibilizados pelo Cadastro de Escolas. Todos os e-mails possuíam um link em comum para um formulário online em que suas respostas seriam salvas para posterior análise. Nenhuma das perguntas do questionário era obrigatória, com a intenção de adquirir o maior número de respostas possíveis, mesmo que incompletas. A cada semana de espera, durante três semanas, foram realizadas ligações para as escolas que ainda não haviam respondido o formulário, com o objetivo de questioná-las sobre o recebimento do e-mail e lembrá-las da pesquisa. Em conjunto com as ligações, o link para o questionário foi reenviado para todas as escolas que não haviam respondido. Durante a primeira semana da coleta de dados, apenas duas escolas responderam o questionário. Após a primeira leva de ligações, mais sete escolas completaram o questionário, seguidas de mais seis durante a terceira e última semana. Das cinquenta e sete escolas contatadas, vinte e duas não foram alcançadas, três se recusaram a participar da pesquisa, e dezessete prometeram responder o questionário, mas nunca o fizeram. Surpreendentemente, uma das estaduais respondeu o

questionário durante a segunda leva de e-mails, apesar de não ter sido possível alcançá-la por telefone. Enquanto isso, outra escola da rede privada foi contatada por telefone três vezes, e em cada uma das vezes forneceu um endereço de e-mail diferente para contato, mas, ao final, não respondeu ao questionário. Apesar de nenhuma das perguntas do questionário ser obrigatória para seu envio, todos os questionários respondidos foram preenchidos por inteiro. Ao fim do período de coleta, o tamanho final da amostra (n) alcançado pela amostra de e-mails foi de quinze escolas. Entre as quinze escolas, nove eram de dependência estadual, quatro de dependência municipal e duas de dependência privada. Devido a um erro por parte do respondente de uma das escolas da rede municipal, onde foi escrito um valor inválido para o número de docentes de ensino fundamental, o tamanho final da amostra (e-mail) para a variável QT_DOC_FUND foi de quatorze escolas.

5.4 CÁLCULOS

Considerando a amostragem estratificada, foram calculados a média populacional, a média amostral e o erro padrão da média amostral, assim como a proporção populacional, a proporção amostral e o erro padrão da proporção amostral de ambas amostras. Para as variáveis dicotômicas foram calculadas suas proporções, considerando uma resposta de “Sim” como um sucesso e “Não” como uma falha.

Montgomery (2012) propõe o nível de confiança de 95% como ideal para análises da média populacional, pois oferece um equilíbrio adequado entre a precisão dos resultados e a probabilidade de que o intervalo de confiança contenha o verdadeiro parâmetro populacional. Níveis de confiança mais altos, como 99%, aumentam a largura do intervalo, tornando as estimativas menos precisas. Por outro lado, níveis mais baixos, como 90%, reduzem sua confiabilidade.

Para pequenas amostras com variância desconhecida, Triola (2013) indica a utilização da distribuição t de Student para o cálculo do intervalo de confiança. A distribuição t de student parte do pressuposto de normalidade das variáveis na população. Entretanto, devido ao tempo limitado e o escopo da pesquisa, não foi possível garantir que as variáveis possuíam distribuições com baixa assimetria. Para a distribuição t de Student, os graus de liberdade das amostras foram de 14 e 18, para as amostras por *web* e telefone respectivamente. Para a variável QT_DOC_FUND da amostra de e-mails foi utilizado grau de liberdade 13, devido à sua variável faltante. Da mesma forma, para os cálculos das variáveis categóricas dos dados

coletados por telefone, foram usados 22 graus de liberdade, devido às respostas adicionais das escolas que responderam o questionário de forma incompleta.

Para encontrar os valores críticos foi utilizado uma tabela de distribuição t de Student. Foi localizado a linha correspondente aos graus de liberdade e a coluna correspondente ao nível de confiança desejado, representado como $\alpha/2$, onde $\alpha = 1 - \text{nível de confiança}$. Os valores críticos encontrados foram de 2,145 e 2,101 para os dados de e-mails e telefones, respectivamente. Para a variável QT_DOC_FUND da amostra de e-mails foi encontrado valor crítico de 2,160 e para as variáveis categóricas da amostra de telefones foi encontrado valor crítico de 2,074.

5.5 RESULTADOS

Ao seguir as ideias propostas por Dillman (2017), foi possível perceber uma menor taxa de respostas por parte das pesquisas por e-mail ao se comparar com as pesquisas por telefone, com apenas 26% das escolas contatadas por e-mail tendo respondido o questionário, comparado à taxa de resposta de 40% dos respondentes por telefone. É notável dizer que um número semelhante de escolas foram capazes de ser contatadas para ambas amostras, seja por e-mails ou telefonemas. Isto pode indicar uma taxa de desinteresse por parte dos respondentes de 50% para pesquisas por e-mail e 23% para pesquisas por telefone.

Os cálculos e análises foram feitos por meio de pacotes como *sampling*, desenvolvido por Tillé e Matei (2023), e *survey*, desenvolvido por Lumley (2024) na interface R Studio, do software livre R. Os pacotes contêm um conjunto de funções para o cálculo de estimativas de médias, proporções e variâncias de diferentes planos amostrais.

Os resultados para a coleta por e-mail estão descritos nas tabelas 1 e 2, e os resultados para a coleta por telefone estão descritos nas tabelas 3 e 4.

Tabela 1 – Resultados das médias para a coleta por e-mail

	Média Populacional (\bar{Y})	Média Amostral (\bar{y})	Erro Padrão da Média Amostral ($ep(\bar{y})$)
QT_DOC_INF	2.4878	2.3891	1.4785
QT_DOC_FUND	21.1951	25.447	5.7997
QT_DOC_MED	13.6860	13.414	6.4998

QT_TUR_INF	1.4604	1.5376	0.7163
QT_TUR_FUND	11.2805	14.292	3.1366
QT_TUR_MED	3.9604	5.272	2.8135

Fonte: Elaborada pelo autor (2024).

Tabela 2 – Resultados das proporções para a coleta por e-mail

	Proporção Populacional (P)	Proporção Amostral (\hat{p})	Erro Padrão da Proporção Amostral ($ep(\hat{p})$)
IN_BIBLIOTECA	0.7713	1	0
IN_LABORATORIO _INFORMATICA	0.8018	0.84756	0.07129
IN_INTERNET	0.9878	1	0
IN_COZINHA	0.9848	1	0
IN_PISCINA	0.0335	0.066734	0.06292
IN_AUDITORIO	0.2073	0.44715	0.11828

Fonte: Elaborada pelo autor (2024).

Tabela 3 – Resultados das médias para a coleta por telefone

	Média Populacional (\bar{Y})	Média Amostral (\bar{y})	Erro Padrão da Média Amostral ($ep(\bar{y})$)
QT_DOC_INF	2.4878	1.3603	0.4814
QT_DOC_FUND	21.1951	24.428	3.2711
QT_DOC_MED	13.6860	10.744	2.7545
QT_TUR_INF	1.4604	0.91391	0.3939
QT_TUR_FUND	11.2805	11.013	1.3887
QT_TUR_MED	3.9604	4.6714	1.5547

Fonte: Elaborada pelo autor (2024).

Tabela 4 – Resultados das proporções para a coleta por telefone

	Proporção Populacional (P)	Proporção Amostral (\hat{p})	Erro Padrão da Proporção Amostral ($ep(\hat{p})$)
IN_BIBLIOTECA	0.7713	0.985976	0.01254
IN_LABORATORIO _INFORMATICA	0.8018	0.87031	0.06959
IN_INTERNET	0.9878	1	0
IN_COZINHA	0.9848	0.957927	0.01254
IN_PISCINA	0.0335	0.028049	0.01254
IN_AUDITORIO	0.2073	0.29995	0.09951

Fonte: Elaborada pelo autor (2024).

Por fim, as médias e proporções populacionais foram comparadas aos intervalos de confiança criados a partir das médias e proporções amostrais. Seus resultados são descritos nas tabelas 5 e 6. Os intervalos de confiança podem ser vistos também através dos gráficos 2, 3, 4 e 5, com as linhas pretas representando os intervalos, e as linhas vermelhas representando a média da população.

Tabela 5 – Resultados intervalos de confiança para a coleta por e-mail

	Média ou Proporção Populacional (\bar{Y})(P)	Limite Inferior	Limite Superior
QT_DOC_INF	2.4878	-0.7822825	5.560482
QT_DOC_FUND	21.1951	12.91965	37.97435
QT_DOC_MED	13.6860	-0.528071	27.35607
QT_TUR_INF	1.4604	0.0011365	3.074064
QT_TUR_FUND	11.2805	7.563993	21.02001
QT_TUR_MED	3.9604	-0.7629575	11.30696
IN_BIBLIOTECA	0.7713	1	1

IN_LABORATORIO _INFORMATICA	0.8018	0.674244	1.020876
IN_INTERNET	0.9878	1	1
IN_COZINHA	0.9848	1	1
IN_PISCINA	0.0335	-0.07312	0.206588
IN_AUDITORIO	0.2073	0.175164	0.719136

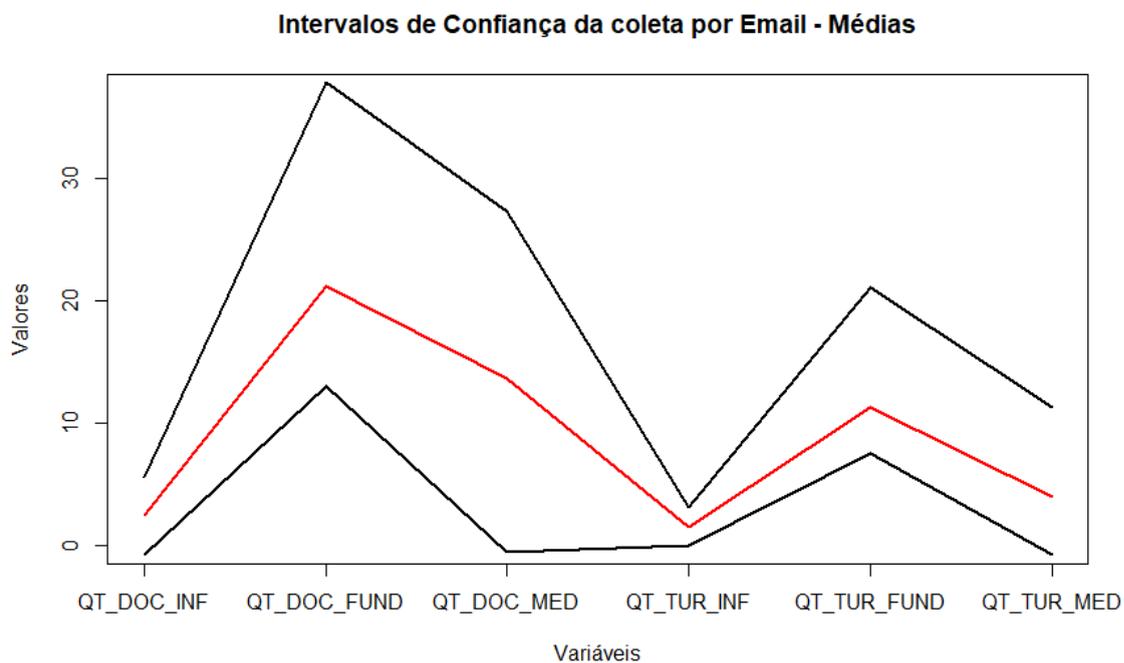
Fonte: do próprio autor (2024)

Tabela 6 – Resultados intervalos de confiança para a coleta por telefone

	Média ou Proporção Populacional (\bar{Y})(P)	Limite Inferior	Limite Superior
QT_DOC_INF	2.4878	0.3488786	2.371721
QT_DOC_FUND	21.1951	17.55542	31.30058
QT_DOC_MED	13.6860	4.956795	16.5312
QT_TUR_INF	1.4604	0.0863261	1.741494
QT_TUR_FUND	11.2805	8.095341	13.93066
QT_TUR_MED	3.9604	1.404975	7.937825
IN_BIBLIOTECA	0.7713	0.9599602	1.011992
IN_LABORATORIO _INFORMATICA	0.8018	0.7259803	1.01464
IN_INTERNET	0.9878	1	1
IN_COZINHA	0.9848	0.9319112	0.9839428
IN_PISCINA	0.0335	0.002033159	0.05406484
IN_AUDITORIO	0.2073	0.09357456	0.5063254

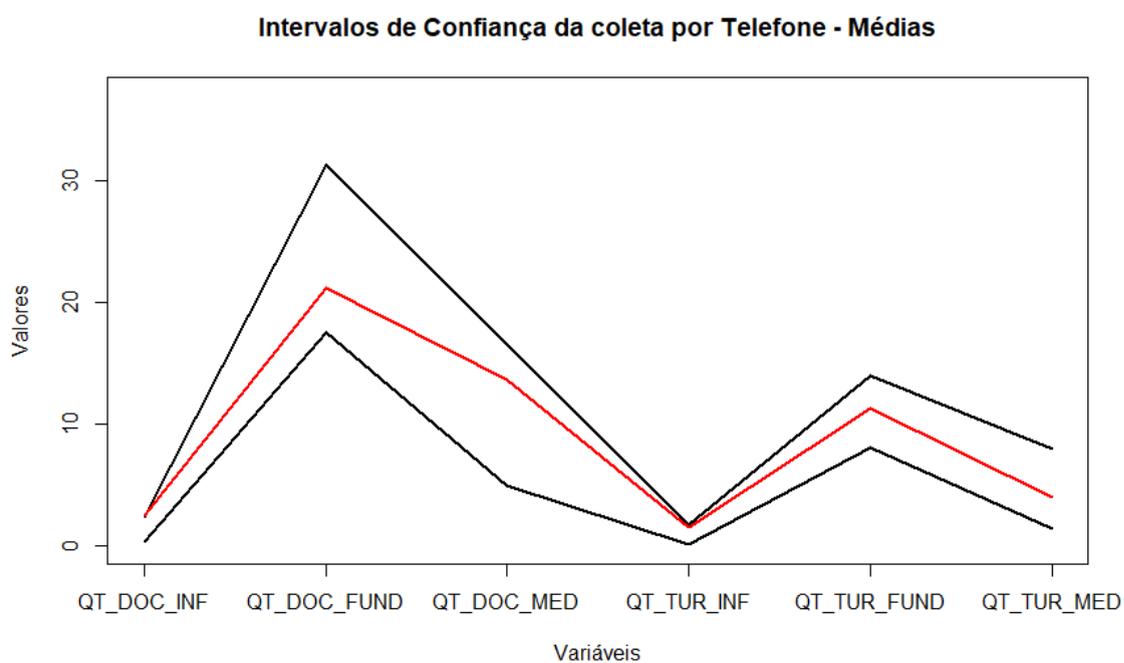
Fonte: do próprio autor (2024)

Gráfico 2 – Intervalos de Confiança da coleta por E-mail - Médias



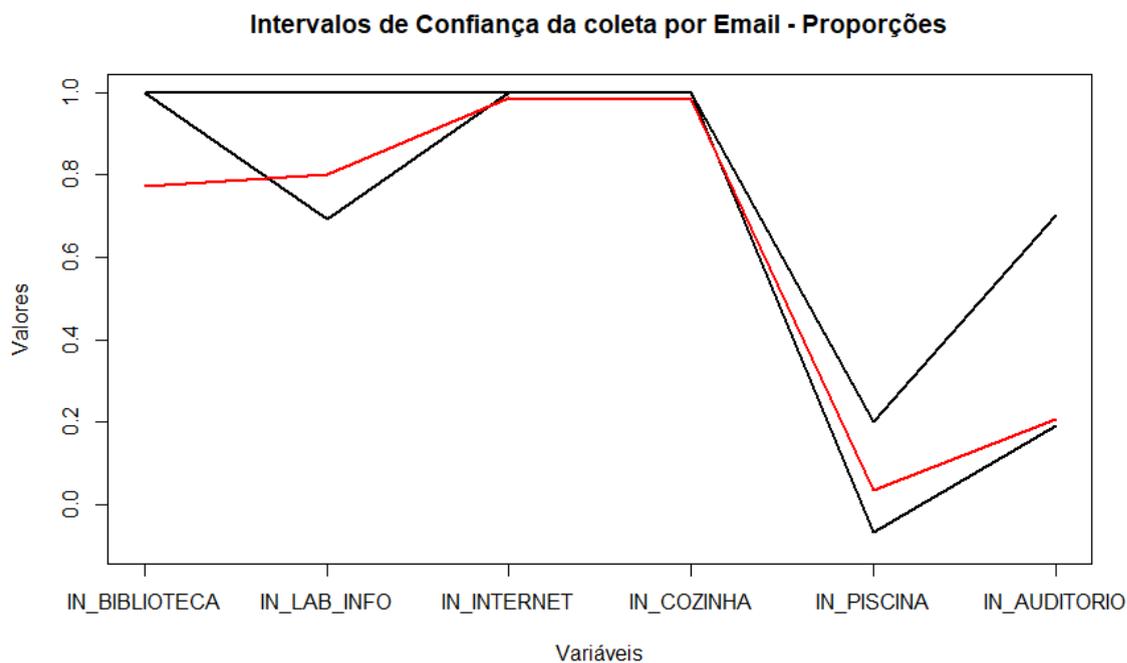
Fonte: do próprio autor (2024)

Gráfico 3 – Intervalos de Confiança da coleta por Telefone - Médias



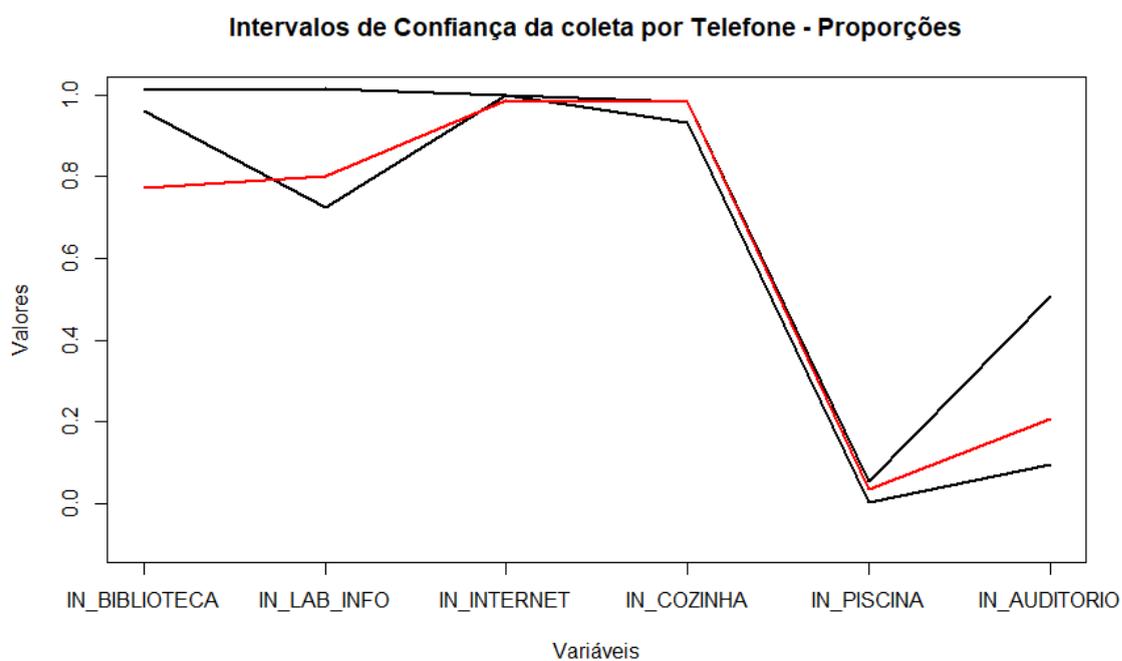
Fonte: do próprio autor (2024)

Gráfico 3 – Intervalos de Confiança da coleta por E-mail - Proporções



Fonte: do próprio autor (2024)

Gráfico 4 – Intervalos de Confiança da coleta por E-mail - Proporções



Fonte: do próprio autor (2024)

Pode ser notado que para os dados coletados por e-mail, apenas três intervalos de confiança não incluíram a média populacional de suas variáveis, sendo estes as variáveis categóricas IN_BIBLIOTECA, IN_INTERNET e IN_COZINHA. Para os dados coletados por telefone, entretanto, quatro intervalos de confiança não incluíram a média populacional de suas variáveis, sendo estes as variáveis QT_DOC_INF, IN_BIBLIOTECA, IN_INTERNET e IN_COZINHA.

Das variáveis com intervalos com ausência da média populacional, três foram similares entre métodos, sendo estas as variáveis categóricas IN_BIBLIOTECA, IN_INTERNET e IN_COZINHA.

6. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Ao comparar as taxas de resposta de ambas amostras, acredita-se que os respondentes da pesquisa feita por e-mail não tiveram o incentivo suficiente para responderem o questionário, comparados aos respondentes da pesquisa por telefone. Estes dados, em conjunto com as taxas de desinteresse, estão de acordo com o que foi proposto por Dillman (2017), onde ele trás a tona como as caixas de entrada de e-mail dos computadores são espaços lotados, onde e-mails não solicitados são mais frequentes, e por consequência mais ignorados, do que ao se comparar com ligações indesejadas. Isto pôde ser confirmado durante as ligações posteriores da pesquisa por e-mails, tal qual uma das mais frequentes justificativas para a não resposta até então foi a falta de percepção ao e-mail pelo respondente. Isto pode ser corroborado adicionalmente pelo que foi dito por Anderson (2024), que aponta como a presença de um entrevistador em tempo real pode garantir que o entrevistado não perca o interesse ou se canse durante o questionário.

Entretanto, é notável a comparação entre a velocidade das respostas dos dois métodos de coleta, assim como o custo em tempo e mão de obra. Um fator a ser apontado é que de todas escolas que responderam o questionário por telefone, 83% dos questionários foram aplicados na primeira semana de contato, enquanto o restante foi adicionado à amostra durante as seguintes semanas de tentativas. Adicionalmente, apenas 79% das escolas responderam o questionário por completo. Enquanto isso, apenas 13% das escolas foram capazes de responder a pesquisa enviada pela primeira leva de e-mails, com um número considerável de escolas adicionais sendo acrescentadas à amostra ao decorrer da coleta de dados. Isto demonstra que posteriores contatos em uma pesquisa por e-mails podem ajudar

significativamente as taxas de resposta, enquanto o mesmo não pode ser dito para pesquisas realizadas singularmente por telefone, na qual apenas o contato inicial já pode ser suficiente para a maior porcentagem das respostas pretendidas. Adicionalmente, a praticidade dos meios de contato também se tornou um fator diferencial entre as pesquisas. Para as pesquisas por e-mail foi necessário um contato mínimo e direto, muitas vezes sendo capaz de alcançar as escolas através de secretários e auxiliares que por sua parte repassaram a mensagem para o respondente pretendido. Enquanto isso, para as pesquisas por telefone, foram necessárias numerosas ligações apenas para alcançar os diretores ou auxiliares que poderiam pessoalmente completar o questionário em um horário em que estivessem disponíveis, o que causou uma coleta de dados mais prolongada e custosa.

Posteriormente, os resultados da análise prática demonstraram uma precisão sutilmente maior para a coleta de dados por e-mail, no que ela apresentou apenas três intervalos de confiança nos quais não incluíram a média populacional, enquanto os dados coletados por telefone apresentaram quatro. Os resultados podem ser interpretados como o fato de que ambos métodos de coleta apresentam valores de precisão dos dados bem próximos, mas que ainda pode haver motivo para a preferência de um dos métodos sobre o outro.

Conclui-se, finalmente, que mesmo que hajam motivos para preferência de um método de contato sobre o outro em relação à qualidade das respostas, sua diferença não chega a ser significativa por si só, e a decisão deve ser feita em cima de um julgamento baseado no custo benefício entre taxas de respostas maiores e taxas de desinteresse menores da pesquisa por telefones contra a agilidade e simplicidade de pesquisas pela *web*.

7. LIMITAÇÕES DO ESTUDO

É de considerável importância notar-se que a pesquisa prática realizada foi feita com uma população de escolas, as quais possuem naturalmente o costume de atender ligações não solicitadas e abrir e-mails de remetentes desconhecidos. Assim como apontado por Dillman (2017), a predisposição de uma população fora destes parâmetros pode variar significativamente, devido à falta de confiança do público em e-mails e telefonemas indesejados. Estes problemas podem ser mitigados caso a organização realizando a pesquisa

seja conhecida ou seja de órgãos governamentais de confiança, mas ainda não garantem os mesmos padrões de resposta vistos nesta pesquisa.

Adicionalmente, devido às limitações de tempo e objetivo deste trabalho, não foi possível assegurar o pressuposto de normalidade das variáveis na população. Devido às consequências que isso pode trazer aos resultados finais dos cálculos, seria válida uma futura pesquisa para expandir neste quesito e garantir simetria nos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

BIEMER, Paul P. Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. *Journal of Official Statistics*, Vol. 17, No. 2, p. 295-320, 2001

BIEMER, Paul P.; MURPHY, Joe; ZIMMER, Stephanie; BERRY, Chip; DENG, Grace; LEWIS, Katie. Using bonus monetary incentives to encourage web response in mixed-mode household surveys. *Journal of Survey Statistics and Methodology*, v. 6, n. 2, p. 240-261, 2018.

Bolfarine, H; Bussab, W. O. (2005) *Elementos de Amostragem*. São Paulo, Edgard Blucher. Brasil. Casa Civil. 90% dos lares brasileiros já têm acesso à internet, aponta pesquisa. 2022. Disponível em: <https://www.gov.br/casacivil/pt-br/assuntos/noticias/2022/setembro/90-dos-lares-brasileiros-ja-tem-acesso-a-internet-no-brasil-aponta-pesquisa>. Acesso em: 1 jun. 2024.

BRASIL. Constituição (1988). *Constituição da República Federativa do Brasil*. Brasília, DF: Senado Federal, 1988.

BUELENS, Bart; VAN DEN BRAKEL, Jan A. Comparing two inferential approaches to handling measurement error in mixed-mode surveys. *Journal of Official Statistics*, v. 33, n. 2, p. 513-531, 2017.

COCHRAN, W. G. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons, 1977.

CETIC.BR. Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros: indicadores. Disponível em: <https://cetic.br/pt/pesquisa/domicilios/indicadores/>. Acesso em: 29 set. 2024.

DILLMAN, Don A. The promise and challenge of pushing respondents to the Web in mixed-mode surveys. *Survey Methodology*, v. 43, n. 1, p. 3-30, 2017.

IBGE. 92,5% domicílios tinham acesso à Internet no Brasil. 2023. Disponível em: <https://educa.ibge.gov.br/jovens/materias-especiais/21581-informacoes-atualizadas-sobre-tecnologias-da-informacao-e-comunicacao.html>. Acesso em: 29 set. 2024.

INEP. *Censo Escolar: Instrumentos de Coleta e Informações Gerais*. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022. Disponível em: <http://portal.mec.gov.br/component/tags/tag/censo-escolar?start=20>. Acesso em: 15 jun. 2024.

ISTOÉ DINHEIRO. Brasil tem 136,3 milhões de acessos à Internet pelo telefone celular para uso pessoal, diz IBGE. 2023. Disponível em: <https://istoedinheiro.com.br/brasil-telefone-celular-uso-pessoal-ibge/>. Acesso em: 14 jun. 2024.

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

Lumley, T. *survey: analysis of complex survey samples*. R package version 4.4-2, 2024

Matei, A. `sampling: survey sampling`. R package version 2.10, 2023

MEC. Ministério da Educação. Fundamentos e Implementação do FUNDEB. Brasília, DF: Ministério da Educação, 2023. Disponível em: <http://portal.mec.gov.br/fundeb>. Acesso em: 15 jun. 2024.

Mininel, C. QuestionPro. 4 vantagens de realizar pesquisa online. 2021. Disponível em: <https://www.questionpro.com/blog/pt-br/4-vantagens-de-realizar-pesquisa-online/>. Acesso em: 14 jun. 2024.

MONTGOMERY, D. C.; RUNGER, G. C. Estatística Aplicada e Probabilidade para Engenheiros. 5. ed. Rio de Janeiro: LTC, 2012.

ROSS, S. M. Introdução à Probabilidade e Estatística para Engenheiros e Cientistas. 5. ed. Rio de Janeiro: Elsevier, 2013.

SCHEAFFER, Richard L.; MENDENHALL III, William; OTT, Lyman. Elementary Survey Sampling. 7. ed. Boston: Cengage Learning, 2011.

SCHOUTEN, Barry; BRAKEL, Jan van den; GIESEN, Deirdre; LUITEN, Annemieke; MEERTENS, Vivian. Mixed mode official surveys. Current status and near future. *The Survey Statistician*, v. 87, p. 12-26, 2023

SCHUTTEN, G; Chan, C. `readODS: read ODS files into r`. R package version 1.6.7, 2018

The Farnsworth Group. "Telephone Survey Advantages and Disadvantages in Manufacturing Market Research". 2024. Disponível em: <https://www.thefarnsworthgroup.com/blog/telephone-survey-market-research> . Acesso em: 27 de maio de 2024.

TRIOLA, M. F. Introdução à Estatística. 12. ed. Rio de Janeiro: LTC, 2013.

ANEXO A - CÓDIGOS PARA OS DADOS DA POPULAÇÃO

leitura dos dados brutos do censo inep

```
censo23 <- read_ods("dados_censo.ods")
```

filtragem dos dados do censo inep para a zona da mata

```
vinteteste <- censo23[censo23$NO_MESORREGIAO == "Zona da Mata", ]
```

variáveis

```
colunas <- c("NO_ENTIDADE", "NO_MUNICIPIO", "CO_ENTIDADE", "NU_DDD",
"NU_TELEFONE", "TP_DEPENDENCIA", "NO_MICRORREGIAO", "QT_DOC_INF",
"QT_DOC_FUND", "QT_DOC_MED", "QT_TUR_INF", "QT_TUR_FUND",
"QT_TUR_MED", "IN_BIBLIOTECA", "IN_LABORATORIO_INFORMATICA",
"IN_INTERNET", "IN_COZINHA", "IN_PISCINA", "IN_AUDITORIO")
```

```
vinteteste <- vinteteste[colunas]
```

retirada das escolas de dependência federal

```
vintao <- subset(vinteteste, TP_DEPENDENCIA != 1)
```

transformação das variáveis para melhor visualização

```
vintao$TP_DEPENDENCIA <- factor(vintao$TP_DEPENDENCIA, levels = c(2, 3, 4), labels
= c("Estadual", "Municipal", "Privada"))
```

```
vintao$TP_DEPENDENCIA <- as.character(vintao$TP_DEPENDENCIA)
```

retirada dos dados incompletos

```
doisfiltro <- vintao[complete.cases(vintao[, c("NO_ENTIDADE", "NO_MUNICIPIO",
"CO_ENTIDADE", "NU_DDD", "NU_TELEFONE", "TP_DEPENDENCIA",
"NO_MICRORREGIAO", "QT_DOC_INF", "QT_DOC_FUND", "QT_DOC_MED",
"QT_TUR_INF", "QT_TUR_FUND", "QT_TUR_MED", "IN_BIBLIOTECA",
"IN_LABORATORIO_INFORMATICA", "IN_INTERNET", "IN_COZINHA",
"IN_PISCINA", "IN_AUDITORIO"))], ]
```

leitura dos dados brutos secretaria da educação de minas

```
caminho2 <- "C:/Users/rafae/OneDrive/Desktop/Rafael/tcc/cadastro_lup.ods"
```

```
dadoslup <- read_ods(caminho2, sheet = 1, range = "A9:Q16774")
```

padronização das células de CO_ENTIDADE para junção dos bancos de dado

```
filtro$CO_ENTIDADE <- substr(doisfiltro4$CO_ENTIDADE,
nchar(doisfiltro4$CO_ENTIDADE)-4, nchar(doisfiltro4$CO_ENTIDADE))
```

```
lup$Codigo_Escola <- substr(dadoslup$Codigo_Escola, nchar(dadoslup$Codigo_Escola)-4,
nchar(dadoslup$Codigo_Escola))
```

fusão dos bancos de dado do censo inep da secretaria da educação de minas

```
df_merged <- merge(filtro, lup[, c("CO_ENTIDADE", "email", "Telefone_Fixo")], by =
"CO_ENTIDADE", all.x = TRUE)
```

```
df_merged2 <- df_merged[!is.na(df_merged$email), ]
```

```
df_merged2 <- df_merged2[, -ncol(df_merged2)]
```

organização da população pelos estratos

```
zonamata2 <- zonamata[order(zonamata$TP_DEPENDENCIA), ]
```

seleção das células sorteadas para a amostra após o sorteio

```
df_estadual <- df_merged2[df_merged2$TP_DEPENDENCIA == "Estadual", ]
```

```
df_municipal <- df_merged2[df_merged2$TP_DEPENDENCIA == "Municipal", ]
```

```
df_privada <- df_merged2[df_merged2$TP_DEPENDENCIA == "Privada", ]
```

```
df_privada_email <- df_privada[c(17, 12, 2, 7, 19, 1, 11, 6, 8, 9, 10, 18),]
```

```
df_privada_tel <- df_privada[-c(17, 12, 2, 7, 19, 1, 11, 6, 8, 9, 10, 18),]
```

```
df_estadual2 <- df_estadual[c(14, 193, 117, 11, 181, 146, 197, 23, 177, 30, 139, 16, 82, 118,  
64, 128, 100, 179, 84, 40, 68, 9, 138, 107, 173, 44, 104, 180, 158, 143, 154, 114, 116, 137,  
172, 65, 77, 61, 88, 157, 18, 97, 156, 194, 92, 108),]
```

```
df_estadual_email <- df_estadual2[c(1:23),]
```

```
df_estadual_tel <- df_estadual2[-c(1:23),]
```

```
df_municipal2 <- df_municipal[c(100, 68, 8, 76, 98, 30, 16, 19, 79, 64, 80, 83, 93, 35, 42, 94,  
57, 12, 6, 62, 47, 9, 15, 92, 54, 48, 71, 73, 18, 81, 27, 4, 89, 108, 66, 84, 13, 3, 91, 107, 43,  
75, 61, 29, 34),]
```

```
df_municipal_email <- df_municipal2[c(1:22),]
```

```
df_municipal_tel <- df_municipal2[-c(1:22),]
```

ANEXO B - CÓDIGOS PARA OS DADOS DA COLETA POR E-MAIL

leitura dos dados

```
data <- read_ods("amostra_email.ods")
```

variáveis

```
colnames(data5) <-
```

```
c("data_hora","NO_ENTIDADE","TP_DEPENDENCIA","QT_DOC_INF","QT_DOC_FUND",
"QT_DOC_MED","QT_TUR_INF","QT_TUR_FUND","QT_TUR_MED","IN_BIBLIOTECA",
"IN_LABORATORIO_INFORMATICA","IN_INTERNET","IN_COZINHA","IN_PISCINA",
"IN_AUDITORIO")
```

organização da amostra pelos estratos

```
data2 <- data2[order(data2$TP_DEPENDENCIA), ]
```

estratificação teórica para o cálculo da probabilidade da amostra

```
IAESs=sampling::strata(zonamata2, stratanames=c("TP_DEPENDENCIA"),
c(9,4,2),method=c("srswor"))
```

correção para população finita

```
fpc_email=rep(c(197,108,23),c(9,4,2))
```

especificando o estilo de amostragem utilizado

```
Plano=svydesign(~1, strata=~TP_DEPENDENCIA, data = data2, probs=~IAESs$Prob,
fpc=~fpc_email)
```

médias amostrais e erros padrão das médias amostrais

```
svymean(~QT_DOC_INF,Plano)
```

```
svymean(~QT_DOC_FUND,Plano)
```

```
svymean(~QT_DOC_MED,Plano)
```

svymean(~QT_TUR_INF,Plano)

svymean(~QT_TUR_FUND,Plano)

svymean(~QT_TUR_MED,Plano)

svymean(~IN_BIBLIOTECA,Plano)

svymean(~IN_LABORATORIO_INFORMATICA,Plano)

svymean(~IN_INTERNET,Plano)

svymean(~IN_COZINHA,Plano)

svymean(~IN_PISCINA,Plano)

svymean(~IN_AUDITORIO,Plano)

erro padrão das proporções amostrais para variáveis categóricas

IN_BIBLIOTECA

$\sqrt{(((197/328)^2)*((1*0)/9))+(((108/328)^2)*((1*0)/4))+(((23/328)^2)*((1*0)/2))}$

IN_LABORATORIO_INFORMATICA

$\sqrt{(((197/328)^2)*((1*0)/9))+(((108/328)^2)*((0.75*0.25)/4))+(((23/328)^2)*((1*0)/2))}$

IN_INTERNET

$\sqrt{(((197/328)^2)*((1*0)/9))+(((108/328)^2)*((1*0)/4))+(((23/328)^2)*((1*0)/2))}$

IN_COZINHA

$\sqrt{(((197/328)^2)*((1*0)/9))+(((108/328)^2)*((1*0)/4))+(((23/328)^2)*((1*0)/2))}$

IN_PISCINA

$\sqrt{(((197/328)^2)*(((1/9)*(8/9))/9))+(((108/328)^2)*((0*1)/4))+(((23/328)^2)*((0*1)/2))}$

IN_AUDITORIO

$\text{sqrt}(\left(\left(\frac{197}{328}\right)^2 \cdot \left(\frac{3}{9} \cdot \frac{6}{9}\right) / 9\right) + \left(\left(\frac{108}{328}\right)^2 \cdot \frac{0.75 \cdot 0.25}{4}\right) + \left(\left(\frac{23}{328}\right)^2 \cdot \frac{0 \cdot 1}{2}\right))$

ANEXO C - CÓDIGOS PARA OS DADOS DA COLETA POR TELEFONE

leitura dos dados

```
tel <- read_ods("amostra_tel.ods")
```

variáveis

```
colnames(tel) <- c("data_hora", "NO_ENTIDADE", "TP_DEPENDENCIA",  
"QT_DOC_INF", "QT_DOC_FUND", "QT_DOC_MED", "QT_TUR_INF",  
"QT_TUR_FUND", "QT_TUR_MED")
```

organização da amostra pelos estratos

```
tel <- tel[order(tel$TP_DEPENDENCIA), ]
```

estratificação teórica para o cálculo da probabilidade da amostra para os dados numéricos

```
IAESs=sampling::strata(zonamata2, stratanames=c("TP_DEPENDENCIA"), c(9,7,3),  
method=c("srswor"))
```

correção para população finita

```
fpc_tel=rep(c(197,108,23),c(9,7,3))
```

especificando o estilo de amostragem utilizado

```
Plano=svydesign(~1, strata=~TP_DEPENDENCIA, data = tel, probs=~IAESs$Prob,  
fpc=~fpc_tel)
```

médias amostrais e erros padrão das médias amostrais para os dados numéricos

```
svymean(~QT_DOC_INF,Plano)
```

```
svymean(~QT_DOC_FUND,Plano)
```

```
svymean(~QT_DOC_MED,Plano)
```

```
svymean(~QT_TUR_INF,Plano)
```

```
svymean(~QT_TUR_FUND,Plano)
```

```
svymean(~QT_TUR_MED,Plano)
```

leitura dos dados categóricos

```
tel2 <- read_ods("amostra_tel2.ods")
```

variáveis

```
colnames(tel2) <- c("data_hora", "NO_ENTIDADE", "TP_DEPENDENCIA",  
"IN_BIBLIOTECA", "IN_LABORATORIO_INFORMATICA", "IN_INTERNET",  
"IN_COZINHA", "IN_PISCINA", "IN_AUDITORIO")
```

organização da amostra pelos estratos

```
tel2 <- tel2[order(tel2$TP_DEPENDENCIA), ]
```

estratificação teórica para o cálculo da probabilidade da amostra para os dados categóricos

```
IAESs=sampling::strata(zonamata2, stratanames=c("TP_DEPENDENCIA"),  
c(11,7,5),method=c("srswor"))
```

correção para população finita

```
fpc_tel2=rep(c(197,108,23),c(11,7,5))
```

especificando o estilo de amostragem utilizado

```
Plano=svydesign(~1, strata=~TP_DEPENDENCIA, data = tel2, probs=~IAESs$Prob,  
fpc=~fpc_tel2)
```

médias amostrais e erros padrão das médias amostrais para os dados numéricos

IN_BIBLIOTECA

```
svymean(~IN_BIBLIOTECA,Plano)[2]
```

```
sqrt((((197/328)^2)*((1*0)/11))+(((108/328)^2)*((1*0)/7))+(((23/328)^2)*((0.8*0.2)/5)))
```

IN_LABORATORIO_INFORMATICA

```
svymean(~IN_LABORATORIO_INFORMATICA,Plano)[2]
sqrt((((197/328)^2)*(((10/11)*(1/11))/11))+(((108/328)^2)*(((6/7)*(1/7))/7))+(((23/328)^2)*((0.6*0.4)/5)))
```

IN_INTERNET

```
svymean(~IN_INTERNET,Plano)[2]
sqrt((((197/328)^2)*((1*0)/11))+(((108/328)^2)*((1*0)/7))+(((23/328)^2)*((1*0)/5)))
```

IN_COZINHA

```
svymean(~IN_COZINHA,Plano)[2]
sqrt((((197/328)^2)*((1*0)/11))+(((108/328)^2)*((1*0)/7))+(((23/328)^2)*((0.4*0.6)/5)))
```

IN_PISCINA

```
svymean(~IN_PISCINA,Plano)[2]
sqrt((((197/328)^2)*((0*1)/11))+(((108/328)^2)*((0*1)/7))+(((23/328)^2)*((0.4*0.6)/5)))
```

IN_AUDITORIO

```
svymean(~IN_AUDITORIO,Plano)[2]
sqrt((((197/328)^2)*(((3/11)*(8/11))/11))+(((108/328)^2)*(((2/7)*(5/7))/7))+(((23/328)^2)*((0.6*0.4)/5)))
```