Amostragem: Teoria e Prática Usando R

true true true

02 de novembro de 2020

Sumário

Bem-vindo

Este é um livro escrito para apoiar a aprendizagem de **Amostragem**. Nosso objetivo principal é orientar um leitor no caminho para aprender os conceitos, as principais ideias e a usar as ferramentas de amostragem para resolver problemas.

Nossa escolha de temas a incluir no livro foi guiada, em grande parte, por nossa experiência com a coleção de pesquisas por amostragem do IBGE, onde trabalhamos por vários anos, e nossa atuação como professores na graduação e na pós graduação da ENCE. Também reflete nossa perspectiva quanto ao melhor caminho para aprender a trabalhar com **Amostragem**.

Nossa abordagem não busca ser exaustiva e, por esse motivo, são poucas as provas que incluímos dos resultados aqui discutidos. Fizemos a escolha deliberada de não apresentar demonstrações da maioria dos resultados relacionados a valor esperado e variância de estimadores. Essas demonstrações podem intimidar e afastar alguns e, além disso, cremos que estão bem cobertas em diversos outros livros sobre o tema. Aos leitores interessados em verificar os resultados, recomendamos a consulta ao excelente livro de Särndal et al. (1992) ou então às muitas referências incluídas ao longo do texto.

Em contraste, escolhemos enfatizar a apresentação de exemplos e de ferramentas computacionais, algo que não tem cobertura tão ampla na literatura sobre **Amostragem**. Nesse contexto, optamos também por enfatizar o uso de comandos ou recursos básicos do R, em lugar de pacotes mais avançados que estão disponíveis.

O livro está organizado em treze capítulos, nominados a seguir:

1) Introdução
2) Conceitos e Cadastros
3) Visão Geral da Amostragem e Estimação
4) Amostragem Aleatória Simples

5) Estimação de Proporções

6) Estimação de Razões e Funções de Totais
7) Estimação para Domínios de Estudo
8) Amostragem Sistemática Simples
9) Outros Métodos de Amostragem com Equiprobabilidade
10) Amostragem com Probabilidades Proporcionais ao Tamanho

11) Amostragem Estratificada
12) Amostragem Conglomerada
13) Estimadores de Calibração

Cada um dos capítulos é autocontido e vários deles podem ser omitidos num primeiro curso. Com exceção do Capítulo **??**, o material do livro pode ser coberto num curso com cerca de 45 horas de duração, como ministrado várias vezes na pós-graduação da ENCE. Caso necessário, algum(ns) dos Capítulos **??**, **??** ou **??** pode(m) ser suprimidos ou separados para estudo individual. Os Capítulos **??** e **??** podem ser omitidos sem prejuízo da sequência. O conteúdo central do livro é formado por todos os capítulos não citados neste parágrafo. Tal conteúdo formaria, a nosso ver, o mínimo para cobertura num primeiro curso, no nível de graduação, sobre **Amostragem**.

Nossa opção ao escolher essa forma de publicação (livro em formato de hipertexto, hospedado na internet) se deve a dois fatores principais: primeiro, não pretendemos comercializar o livro e, sim, torná-lo de acesso livre e aberto,

como é a filosofia do software que usamos para sua elaboração e produção (R + RStudio + R Markdown + Github); segundo, essa forma de publicação permitirá atualizações mais rápidas e frequentes do conteúdo, o que favorece a correção de erros, revisões do texto, inclusão de exemplos ou tópicos novos etc. Esperamos que essa escolha não afaste os leitores que ainda gostam de livros em papel, como nós...

O leitor de qualquer livro precisa reconhecer que não é possível começar do zero: é preciso contar com conhecimento prévio de algumas ideias e conceitos básicos essenciais à compreensão do material tratado. Nossa abordagem pressupõe que o leitor está familiarizado com um curso básico de introdução à probabilidade e à inferência estatística, no nível tratado, por exemplo, em Magalhães e Lima (2004) e Magalhães (2006).

Capítulo 1

Conceitos e Cadastros

1.1 Fontes de dados e tipos de pesquisas

Uma importante decisão em qualquer estudo ou projeto de pesquisa diz respeito ao levantamento das *fontes de dados* com potencial para atender às necessidades de informações de interesse. Um primeiro critério de classificação distingue as *fontes primárias*, cujos dados ainda não foram coletados, das *fontes secundárias*, cujos dados já foram coletados, possivelmente com outro(s) propósito(s), e estão disponíveis ou poderiam ser obtidos para uso imediato.

No caso da *fonte primária*, a obtenção dos dados pode ser feita através de um *estudo de caso*, de uma *pesquisa* (*survey*) ou de um *experimento*. Para uma discussão mais extensa a esse respeito, ver, por exemplo, o excelente Capítulo 1 de Wild e Seber (2004). Neste livro, nosso foco será sobre fontes primárias do tipo *pesquisa*.

Por definição, *pesquisa* é uma *operação estatística* de coleta de informações sobre características de interesse de unidades de uma população, usando conceitos, métodos e procedimentos bem definidos, de modo que permita a compilação dessas informações numa forma resumida útil.

Dependendo da amplitude da coleta dos dados, há dois tipos de *pesquisa*: censos e pesquisas amostrais. Um *censo* é uma pesquisa baseada numa enumeração exaustiva das unidades componentes de uma população, realizada com o propósito de coletar informações sobre aspectos relevantes dessa população. Num *censo*, a intenção é ter dados referentes a todas as unidades da população. Alternativamente, quando as informações vão ser coletadas somente para um subconjunto selecionado das unidades da população, se diz então que a *pesquisa* é *por amostragem* ou *amostral*.

De acordo com S. K. Thompson (2012) tem-se a seguinte definição: "*Amostragem* consiste em selecionar parte de uma população para observar, de modo que seja possível estimar alguma coisa sobre toda a população."

Neste livro, o principal tipo de pesquisa a ser considerado é a *Pesquisa Amostral*, que busca conhecer a população com base numa amostra. Mas algumas vezes fazemos referência a censos que poderiam, teoricamente, ser conduzidos nas mesmas populações das quais tratamos de tirar amostras para fazer inferência.

1.2 Alguns conceitos fundamentais

Antes de avançar na apresentação da teoria relevante para *pesquisas por amostragem*, vamos introduzir vários conceitos relevantes e necessários para sustentar a discussão posterior.

Unidade é um único indivíduo, entidade ou objeto a ser medido ou observado na pesquisa.

População é o conjunto de todas as unidades para as quais se deseja fazer inferência.

População alvo é a parte da *População* para a qual se gostaria de obter informação. Seja um estudo sobre a fecundidade da *População* de uma determinada área geográfica. A *População Alvo* neste caso são apenas as mulheres em idade fértil.

População de pesquisa é a população a ser efetivamente coberta pela pesquisa. Algumas unidades ou grupos de unidades podem ser inacessíveis ou com custo-benefício que não se mostre interessante para a pesquisa. O Exemplo

1.4 mostra as definições da *população alvo* e da *população de pesquisa* para a PNAD 2003, explicando suas diferenças.

Quando ocorrer diferença entre a *população alvo* e a *população de pesquisa*, o pesquisador responsável deve procurar medir essa diferença e tratar de revelar com clareza que partes da *população alvo* ficaram de fora da *população de pesquisa*, de modo a permitir que os usuários dos resultados da pesquisa possam avaliar a relevância e aderência dos resultados da pesquisa aos seus objetivos. A diferença entre a *população de pesquisa* e a *população alvo* dá origem aos chamados *erros de cobertura*. Pesquisas de boa qualidade adotam estratégias para eliminar ou minimizar tais erros.

Amostra é o conjunto de unidades que selecionamos da *população de pesquisa* para medir ou observar através da pesquisa.

Amostra efetiva é o conjunto de unidades que selecionamos da *população de pesquisa* e para as quais conseguimos de fato obter / medir / observar as variáveis de interesse através da pesquisa.

Cadastro é uma lista contendo a identificação das unidades que compõem a *população de pesquisa*, de onde a amostra é selecionada. Algumas pesquisas podem utilizar mais de um cadastro para selecionar a correspondente amostra.

Unidade de referência ou *unidade de investigação* ou *unidade de observação* é uma unidade (componente da população) sobre a qual são obtidas as informações de interesse da pesquisa.

Unidade informante ou *unidade de informação* é uma unidade que fornece a informação de interesse sobre uma ou mais unidades de referência.

Unidade de análise é uma unidade à qual a análise e inferência são dirigidas.

Unidade de amostragem é uma unidade que pode ser selecionada para a amostra numa das etapas do processo de amostragem.

Domínio de análise / interesse é um conjunto de unidades de análise especificado como de interesse para fins de sumarização de dados, tabulação, inferência ou análise.

Para dar concretude aos vários conceitos aqui introduzidos, vamos considerar alguns exemplos de situações de pesquisa, buscando destacar, em cada uma delas, a aplicação de alguns dos conceitos apresentados.

Exemplo 1.1. Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua

A PNAD Contínua é hoje a principal pesquisa domiciliar realizada pelo IBGE para acompanhamento das condições de vida da população brasileira. Conforme o IBGE "O principal objetivo é produzir informações contínuas sobre a inserção da população no mercado de trabalho e de características tais como idade, sexo e nível de instrução, bem como permitir o estudo do desenvolvimento socioeconômico do País através da produção de dados anuais sobre outras formas de trabalho, trabalho infantil, migração, entre outros temas." Maiores informações sobre a pesquisa e seus métodos podem ser encontradas em Freitas e Antonaci (2014) e também no endereço: https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html?edicao=20106&t=conceitos-e-metodos.

A *população alvo* da pesquisa é "... constituída por todas as pessoas moradoras em domicílios particulares permanentes da área de abrangência da pesquisa." Segundo Freitas e Antonaci (2014), a área de abrangência geográfica da pesquisa é definida como "... todo o território nacional, dividido nos setores censitários da Base Operacional Geográfica de 2010, excluídas áreas com características especiais classificadas pelo IBGE como setores censitários de: aldeias indígenas, quartéis, bases militares, alojamentos, acampamentos, embarcações, penitenciárias, colônias penais, presídios, cadeias, asilos, orfanatos, conventos, hospitais e agrovilas de projetos de assentamentos rurais. ... também foram excluídos os setores censitários localizados em Terras Indígenas."

Não há na documentação da pesquisa indicação de que a *população de pesquisa* é distinta da *população alvo*. Nessa pesquisa, ficam de fora da *população alvo* moradores institucionalizados, isto é, moradores em instituições tais como hotéis e pensões, abrigos e asilos, instalações militares, hospitais, presídios, etc. Também ficam de fora moradores em domicílios improvisados (por exemplo, acampamentos ou áreas de ocupação precária) e pessoas em situação de rua, isto é, que não residem em domicílios de qualquer tipo ou em instituições. Fosse o conjunto da população residente no Brasil declarado como *população alvo*, a *população de pesquisa* da PNAD Contínua teria um déficit de cobertura devido à exclusão destas partes da população.

Nessa pesquisa, as pessoas são definidas como *unidades de referência*, *unidades informantes* e *unidades de análise*. Apesar disso, há duas perguntas no questionário da pesquisa que se referem à habitação (domicílio) - uma é a condição de ocupação do domicílio e, para os domicílios alugados, é levantado o valor do aluguel mensal pago.

Logo seria correto identificar que os domicílios também são *unidades de referência*. Além disso, os domicílios são *unidades de amostragem* pois são objeto de sorteio na última etapa do plano amostral empregado na pesquisa e, também são *unidades de análise* pois são construídas variáveis em nível de domicílio que fazem parte da tabulação da pesquisa, tais como o rendimento domiciliar per capita. Seria também necessário definir uma *população de pesquisa* correspondente a esse conjunto de unidades adicionais, que poderia ser caracterizada como "o conjunto de todos os domicílios particulares permanentes da área de abrangência da pesquisa."

Os níveis de divulgação da pesquisa (*domínios de interesse*) incluem os seguintes grupos definidos em função da localização geográfica dos domicílios: Brasil (1); Grandes Regiões (5); Unidades da Federação (27); Regiões Metropolitanas que contêm Municípios das Capitais (20); Municípios das Capitais (27); e a Região Integrada de Desenvolvimento da Grande Teresina (1). Os números entre parênteses se referem à contagem de domínios definidos em cada uma das situações. Como se pode verificar, ao todo há 81 domínios de interesse definidos com base na localização geográfica que precisam ser contemplados na apresentação de resultados da PNAD Contínua. Vale também notar que há domínios de interesse sobrepostos, tais como as Capitais e as Regiões Metropolitanas que as contêm, para citar um exemplo.

A definição dos *domínios de interesse* foi uma das etapas cruciais para o planejamento da pesquisa, pois condicionou de forma importante a determinação do tamanho e também a alocação e a distribuição espacial da amostra, entre outros aspectos da metodologia. Conforme IBGE (2014), "o tamanho da amostra da PNAD Contínua foi calculado como o necessário para estimar o total de pessoas desocupadas de 14 anos ou mais de idade ... com um nível de precisão pré-determinado ... para cada uma das Unidades da Federação, por ciclo de acumulação trimestral". Ao final, a amostra total resultante deveria ser de cerca de 211.000 domicílios distribuídos em cerca de 15.100 setores censitários, a cada trimestre. Em cada trimestre, a amostra de setores é distribuída para coleta ao longo das 12 semanas, resultando na coleta de cerca de 1.258 setores por semana.

Verifica-se da descrição acima que a determinação dos tamanhos de amostra privilegiou os domínios de interesse definidos pelas *Unidades da Federação*, não tendo sido estabelecidos tamanhos de amostra capazes de dar precisão controlada para indicadores dos domínios que correspondem a áreas menores (as capitais ou as regiões metropolitanas, quando consideradas individualmente, por exemplo).

Exemplo 1.2. Pesquisa de Orçamentos Familiares - POF 2008/2009

Segundo o IBGE, a *população alvo* da pesquisa é composta por "Domicílios particulares permanentes ocupados e seus moradores, na área de abrangência da pesquisa nas situações urbana e rural. Foram excuídas as áreas definidas pelo IBGE como sendo quartéis, bases militares, alojamentos, acampamentos, embarcações, penitenciárias, colônias penais, presídios, cadeias, asilos, orfanatos, conventos e hospitais." Ver detalhes no link: https://www.ibge.gov.br/estatisticas/sociais/rendimento-despesa-e-consumo/9050-pesquisa-de-orcamentos-familiares.html?=&t=conceitos-e-metodos

Essa definição da *população alvo* estaria melhor caracterizada como *população de pesquisa*. Ao explicitar exclusões como as citadas, fica implícita a ideia de que a população alvo seria a população definida pela condição afirmativa "Domicílios particulares permanentes ocupados e seus moradores, na área de abrangência da pesquisa nas situações urbana e rural." As exclusões são feitas para explicitar que parte dessa população não será coberta pela pesquisa, por razões tipicamente operacionais e associadas aos custos e prazos.

Ainda segundo o IBGE, "A Pesquisa de Orçamentos Familiares 2008-2009 teve por objetivo fornecer informações sobre a composição dos orçamentos domésticos, a partir da investigação dos hábitos de consumo, da alocação de gastos e da distribuição dos rendimentos, segundo as características dos domicílios e das pessoas. A POF investigou, também, a autopercepção da qualidade de vida e as características do perfil nutricional da população brasileira."

O plano amostral empregado na pesquisa é descrito como "... conglomerado em dois estágios de seleção, com estratificação geográfica e estatística das unidades de primeiro estágio. Os setores correspondem às unidades do primeiro estágio de seleção e os domicílios particulares permanentes, às unidades do segundo estágio." Sendo assim, a pesquisa identificou duas *unidades de amostragem*: setores censitários (selecionados no primeiro estágio de amostragem), e domicílios (selecionados no segundo estágio de amostragem). Uma vez que um domicílio era selecionado para a amostra no segundo estágio, todos os moradores do mesmo deviam ser pesquisados. Note então que as pessoas moradoras não são *unidades de amostragem*, pois não participam de qualquer etapa de sorteio.

Para fins do dimensionamento da amostra, "Foram fixados diferentes coeficientes de variação para estimar com a precisão desejada o total da renda dos responsáveis pelos domicílios, segundo os diferentes domínios de estimação." "Para estimar o total nas Unidades da Federação da Região Norte, foram fixados coeficientes de variação que

variaram de 10% a 15%.”... Ver detalhes em: https://www.ibge.gov.br/estatisticas/sociais/rendimento-despesa-e-consumo/9050-pesquisa-de-orcamentos-familiares.html?=&t=notas-tecnicas.

Exemplo 1.3. *Canadian Labour Force Survey*

Conforme o *Statistics Canada*, agência responsável pela pesquisa: “The target population is the non-institutionalised population 15 years of age and over. The survey is conducted nationwide, in both the provinces and the territories. Excluded from the survey's coverage are: persons living on reserves and other Aboriginal settlements in the provinces; full-time members of the Canadian Armed Forces, the institutionalized population, and households in extremely remote areas with very low population density. These groups together represent an exclusion of less than 2% of the Canadian population aged 15 and over.” Mais detalhes disponíveis em: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3701&lang=en&db=imdb&adm=8&dis=2#a3.

Nesse exemplo também aparece a ideia de uma *população alvo* que seria maior que a *população de pesquisa* de fato considerada na pesquisa e inclusive há uma estimativa do déficit de cobertura representado pelas exclusões indicadas: 2% da população canadense de 15 anos e mais. Também é interessante notar que a pesquisa foca apenas na população de 15 anos e mais.

A citação a esta pesquisa é útil também porque contém uma explicitação mais complexa dos objetivos que devem ser cumpridos em termos da qualidade com que certas estimativas deveriam ser produzidas. O motivo é que há conjuntos de *domínios de interesse* mais variados. A publicação Statistics Canada (2020) explicita os seguintes objetivos quanto à precisão de estimativas para diferentes domínios:

“For monthly estimates of unemployment at the Canada level:

- a CV lower than 2%.

For monthly estimates of unemployment at the provincial level:

- a CV lower than 7% if the unemployment rate is higher than 5%, or a fixed confidence interval width for the unemployment estimates equivalent to that of an unemployment rate of 5% with a CV of 7%.

For three-month moving average estimates at the sub-provincial level:

- a CV lower than 25% for the unemployment estimate at the Economic Region - ER level.
- a CV lower than 15% for the unemployment rate estimate at the Employment Insurance Economic Regions - EIER level if the unemployment rate is higher than 5%, or a fixed confidence interval width equivalent to that of an unemployment rate of 5% with a CV of 15%."

Assim os objetivos especificados para guiar o dimensionamento da amostra incluíram não só níveis diferenciados de precisão para domínios de diferentes tamanhos, como também envolveram a consideração de períodos de acumulação de amostras (a pesquisa é repetida mensalmente) necessários para permitir obter estimativas para os domínios mais detalhados ou desagregados.

Exemplo 1.4. Pesquisa Nacional por Amostra de Domicílios - PNAD 2003

População alvo - População residente no Brasil numa data de referência especificada.

População de pesquisa - População residente no Brasil numa data de referência especificada, excluídos os habitantes de setores rurais da região Norte (menos os do estado do Tocantins).

Nessa pesquisa, conforme as definições acima, havia uma diferença entre a *população alvo* e a *população de pesquisa* consideradas nas PNADs anteriores ao ano de 2004. Nas pesquisas mais antigas, a população residente nos setores rurais da Região Norte (exceção do Tocantins) era excluída. O déficit de cobertura era modesto (estima-se que cerca de apenas 2% da população brasileira residia nos setores excluídos). Entretanto, analistas interessados em comparar resultados de PNADs anteriores a 2004 com os de PNADs mais recentes devem avaliar com cuidado se seria ou não necessário excluir das PNADs mais recentes os dados das partes que eram excluídas das PNADs mais antigas.

Na série de pesquisas PNAD, encerrada em 2015, os *domínios de análise* (principais) eram as Unidades da Federação (27), as Regiões Metropolitanas (9) situadas em torno de capitais, o total do país (1), e os totais das áreas urbana e rural (2). Para detalhes, consulte, por exemplo, IBGE (2004).

Exemplo 1.5. População de pesquisa para a Pesquisa de Economia Informal Urbana - ECINF

Segundo IBGE (2003), trata-se de "Pesquisa por amostragem de domicílios situados em áreas urbanas, onde se busca identificar os trabalhadores por conta própria e empregadores com até 5 empregados que desenvolvam atividades não agrícolas." Além disso, "... pertencem ao informal cada uma das unidades econômicas de propriedade de trabalhadores por conta própria e de empregadores com até 5 empregados, moradores de áreas urbanas, sejam elas a atividade principal de seus proprietários ou atividades secundárias." Ver IBGE (2003), página 16.

Nesta pesquisa a caracterização da população de pesquisa é mais complexa, pois envolve aplicação de perguntas para identificar a situação ocupacional (só trabalhadores por conta própria e empregadores são elegíveis), bem como o porte e a atividade do estabelecimento onde o trabalho é exercido. Pesquisas assim costumam ter a missão de estimar também o tamanho da população de pesquisa.

Exemplo 1.6. Definindo populações de pesquisa

Para melhor ilustrar o conceito da definição da *população de pesquisa*, a Tabela 1.1 apresenta diversos exemplos de definição para populações de pesquisa com as seguintes especificações: as unidades a serem pesquisadas; características definidoras das unidades; localização espacial das unidades; e período de referência considerado. Esta separação de componentes da definição ajuda a identificar quais são os elementos essenciais que devem compor a definição de uma *população de pesquisa*.

Tabela 1.1: Exemplos de definições de populações de pesquisa

Unidades de Referência	Características Definidoras	Localização	Período
Pessoas	Habitando domicílios particulares permanentes	Em Macaé	Durante a semana da pesquisa
Empresas do comércio varejista	Classificados como supermercados	Em Recife	Em 1996
Pessoas	Maiores de 5 anos de idade	Que visitaram o Museu Nacional	Entre 1 de junho e 30 de setembro de 1996
Alunos	Do curso de mestrado da ENCE	No Rio de Janeiro	Primeiro semestre de 2002
Estabelecimentos agropecuários	Produtores de café	No Paraná	No ano de 1998

Este conjunto de exemplos buscou ilustrar os conceitos mais importantes necessários à especificação de populações que serão objeto de pesquisas por amostragem. Como o alvo da inferência a partir da amostra é sempre uma população que não se vai conhecer por inteiro através da pesquisa, é importante que a definição da população de pesquisa seja a mais clara possível, para permitir análises corretas das estimativas que serão produzidas a partir da amostra.

Uma rica fonte de informações sobre as definições conceituais de uma vasta coleção de pesquisas é a base de metadados das pesquisas do IBGE, acessível através de https://metadados.ibge.gov.br/. Outra boa fonte de exemplos de definições conceituais de pesquisas bem apresentadas são as pesquisas realizadas pelo NIC.br: https://cetic.br/pesquisas/.

1.3 Abordagens alternativas para pesquisas por amostragem

Pesquisas por amostragem dependem criticamente da qualidade da amostra selecionada para coleta ou obtenção dos dados. Para cumprir bem seu papel na pesquisa, são características desejáveis da amostra:

a) Permitir generalizar estimativas dela derivadas para o conjunto da população de pesquisa.
b) 'Imparcialidade'.
c) Fornecer estimativas com o menor erro amostral possível, dados os recursos disponíveis (financeiros, tempo, pessoal e outros) e considerando as restrições operacionais.
d) Assegurar a capacidade de medir a precisão das estimativas dela provenientes.

A base para o processo de amostragem e de inferência que consideramos neste livro não é dada por um modelo que se utiliza para representar distribuições geradoras dos valores da população, mas sim por um *modelo de aleatorização*

que decorre da imposição de um método imparcial (ao acaso, por sorteio) de escolha das unidades da amostra. Este modelo depende de três condições simples de satisfazer, em grande número de aplicações:

1) Cada unidade da população de pesquisa deve ter uma probabilidade positiva, conhecida ou calculável, de ser incluída na amostra.
2) A seleção da amostra deve ser feita por mecanismo aleatório que assegure a condição 1 e que permita conhecer ou calcular as probabilidades de inclusão de cada uma das unidades selecionadas.
3) As probabilidades de inclusão das unidades selecionadas devem ser levadas em consideração ao fazer inferência para os parâmetros da população, juntamente com outros aspectos da estrutura do método usado para seleção da amostra.

Procedimentos de amostragem que satisfazem as condições 1 a 3 formam a base da abordagem denominada de *Amostragem Probabilística*, que é a mais usada para a amostragem de populações finitas e para a elaboração de estatísticas oficiais e públicas ao redor do mundo. Essa abordagem fornece as condições para fazer inferências seguras para os parâmetros da população de pesquisa, com margem de erro conhecida e controlada.

Até este ponto do livro, o termo *Amostragem* vinha sendo usado num sentido amplo. Deve ficar claro, porém, que é a *Amostragem Probabilística*, com suas regras objetivas e precisas para a escolha da amostra baseadas na teoria das probabilidades, que torna possível estimar os parâmetros desejados e avaliar a margem de erro dessas estimativas com base no *modelo de aleatorização*. Sob esse modelo, os valores de variáveis de interesse observados para unidades da população são considerados fixos embora desconhecidos, exceto para as unidades que forem selecionadas para a amostra. Toda a incerteza (fonte de variação estocástica) é introduzida pelo processo de seleção da amostra, usando um mecanismo probabilístico bem especificado. Fica entendido, então, que o termo *Amostragem* será doravante usado com o significado implícito de *Amostragem Probabilística*, exceto quando se explicitar outro significado num contexto ou exemplo específico.

Amostras que não satisfazem as condições 1 a 3 mencionadas acima podem não permitir generalizar a inferência para a população como um todo. Exemplos de métodos de amostragem que não satisfazem as condições indicadas incluem: amostras de conveniência, amostras de voluntários, amostras intencionais (de "corte"), amostras por quotas. Muitas pesquisas são realizadas usando amostras extraídas segundo métodos como esses, que não estão amparados pelos resultados teóricos que sustentam a *Amostragem Probabilística*. Uma dificuldade central para tais pesquisas é que não estão disponíveis métodos adequados para estimação pontual ou para a estimação da precisão quando as amostras não são probabilísticas. Algumas dessas pesquisas se valem dos métodos da *Amostragem Probabilística* para justificar a apresentação de 'margens de erro' de suas estimativas, mas essa prática não tem sustentação na teoria vigente, pois tais amostras não satisfazem as condições aqui explicitadas. Um exemplo conhecido é o das pesquisas de intenções de voto realizadas no Brasil a cada nova eleição que, frequentemente, não utilizam métodos de *Amostragem Probabilística*.

Há outras abordagens sólidas para fundamentar a realização de pesquisas por amostragem. Destaca-se, em particular, a abordagem dos *Modelos de Superpopulação* e da *Amostragem Baseada em Modelos*, muito bem descrita em livros tais como Valliant et al. (2000) ou Chambers e Clark (2012). Nestes textos uma perspectiva clássica da inferência é adotada. Há também abordagens equivalentes que adotam uma perspectiva Bayesiana para a inferência. Nos dois casos, a inferência é governada não por um *mecanismo de aleatorização* introduzido pelo pesquisador para a extração da amostra, mas por modelos que especificam distribuições e estruturas de dependência para as observações da população (os chamados *Modelos de Superpopulação*). A liberdade do pesquisador quanto aos métodos para extração de amostras é maior, mas também é maior a dependência dos resultados da sua inferência quanto à validade e adequação das hipóteses feitas quanto aos *Modelos de Superpopulação* especificados.

Apesar da relevância do tema para um pesquisador interessado nos fundamentos da *Amostragem*, neste livro adotamos uma perspectiva mais restritiva: adotamos a abordagem da *Amostragem Probabilística* e não tratamos das suas alternativas mais bem fundamentadas acima mencionadas. Esta limitação tem explicação na origem de nossa prática profissional que é lastreada, em grande parte, pelas aplicações no IBGE e em suas pesquisas. Mundo afora, as instituições produtoras de estatísticas oficiais adotam esta abordagem como padrão e são ainda raras as instâncias em que pesquisas são planejadas e realizadas com suporte em outras abordagens como as citadas acima.

1.4 Planejamento e execução de pesquisas por amostragem

Pesquisas por amostragem, para serem bem feitas, requerem cuidadoso planejamento, dedicada execução e rigorosa avaliação. Os métodos e processos de trabalho necessários para o sucesso da pesquisa já são bem conhecidos e

descritos na literatura especializada - ver, por exemplo, o excelente livro de Backstrom e Hursh-César (1981) ou o mais moderno de Groves et al. (2009).

Para uma visão de conjunto, o *Modelo Genérico do Processo de Produção Estatística - MGPPE* ou *Generic Statistical Business Process Model - GSBPM*, definido pela *United Nations Economic Commission for Europe* - UNECE, é um modelo que descreve de forma abrangente as atividades do processo de produção estatística. Este modelo tem sido utilizado como quadro de referência para nortear a modernização e a melhoria da qualidade da produção estatística em muitos institutos nacionais de estatística, entre eles o IBGE. Uma representação esquemática dessa abordagem é apresentada na Figura 1.1, adaptada do *Generic Statistical Business Process Model: GSBPM: version 5.0* (GSBPM (2013), página 9).

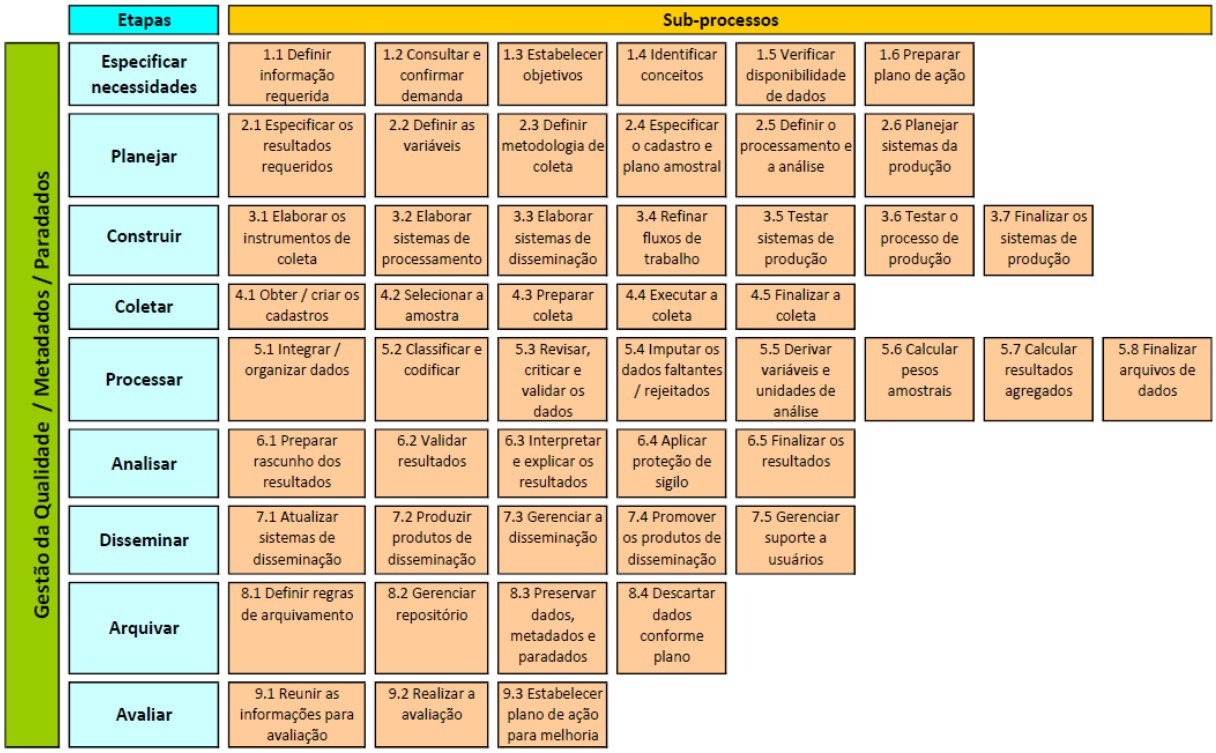


Figura 1.1: Modelo Genérico do Processo de Produção Estatística, adaptado

Conforme indicado neste modelo, o processo de produção de uma pesquisa por amostragem engloba atividades de supervisão e controle em todas as etapas. Como pode ser observado, a gestão da qualidade é transversal ao processo de produção, perpassando todas as etapas definidas no modelo. A seguir, é apresentada uma breve descrição de etapas fundamentais identificadas no processo de produção de uma pesquisa por amostragem.

1.4.1 Especificação das necessidades da pesquisa

Esta etapa envolve os subprocessos que definem os objetivos, os conceitos e a preparação de um plano de ação para a realização da pesquisa.

Quando surge a necessidade de fazer uma pesquisa sobre um certo tema e há uma decisão de alocar recursos para a concretização da pesquisa, deve-se identificar e especificar com clareza e precisão os objetivos específicos da pesquisa. Tal especificação deve incluir a definição da população alvo e de pesquisa, que resultados a pesquisa deve fornecer ou que perguntas principais a pesquisa tem que ajudar a responder, em que prazos e com que recursos poderá ser efetivada. Juntamente com a definição dos recursos, deverá ser especificada a margem de erro tolerável nos principais resultados da pesquisa. O Exemplo 1.1 acima contém vários dos elementos aqui mencionados em relação à PNAD Contínua do IBGE.

O papel do estatístico ou amostrista nesta etapa é o de orientar a tomada de decisões quanto aos objetivos da pesquisa, esclarecendo os interessados quanto às limitações dos métodos estatísticos disponíveis, trabalhando para

garantir a especificação de objetivos viáveis e compatíveis entre si e, também, coerentes com os recursos disponíveis. Cabe também ao estatístico estabelecer projetos ou contratos que garantam a disponibilidade dos recursos necessários para a realização bem sucedida da pesquisa, nos termos dos objetivos especificados.

Nesta etapa, o estatístico deve deixar claro que os resultados de uma pesquisa por amostragem só são aplicáveis à *população de pesquisa* quando a amostragem puder levar em conta todas as unidades pertencentes a essa população, e que quaisquer interesses quanto a resultados e margem de erro de estimativas da pesquisa, se não especificadas a priori, correm o risco de, mais tarde, não poderem ser contemplados. A explicitação de objetivos e recursos é uma das etapas cruciais que devem anteceder a elaboração do *plano amostral*, que vai ser objeto do trabalho do estatístico ou amostrista encarregado de planejar a pesquisa.

1.4.2 Planejamento da pesquisa

Esta etapa envolve a definição das variáveis, a obtenção e avaliação do cadastro, o planejamento da amostra, a metodologia de coleta, a definição do processamento e análise e o fluxo de produção da pesquisa. A questão do tipo de informação a ser coletada deve ser considerada num estágio inicial do planejamento da pesquisa. Somente os dados relevantes para os propósitos da pesquisa devem ser levantados. Se muitas perguntas forem propostas, os informantes perderão o interesse em respondê-las. Por outro lado, deve ser assegurado que nenhum item importante seja esquecido.

Uma regra prática consiste em preparar os leiautes das tabelas que a pesquisa deverá produzir e só depois preparar os instrumentos de coleta dos dados (questionários). Isto ajuda a eliminar informações irrelevantes bem como garantir a inclusão de todos os itens importantes. Uma consideração importante é a possibilidade prática de obtenção da informação: os informantes escolhidos podem não ser capazes de responder a todas as perguntas formuladas. É interessante que essa possibilidade seja, pelo menos, minimizada na fase de planejamento da pesquisa.

O *cadastro* ou *marco de referência* é o instrumento principal que será usado para localizar e identificar as unidades da população de pesquisa. É o cadastro que serve como base para a amostragem de unidades da população, para apoiar a coleta dos dados e também para auxiliar e controlar o processamento dos dados coletados pela pesquisa. Portanto, para poder acessar e cobrir a população de pesquisa definida, é necessário contar com listas das suas unidades, com listas de grupos de unidades populacionais, ou mesmo com mapas ou quaisquer outros materiais que sirvam de guia e permitam localizar e identificar as unidades da população a ser coberta.

Tais listas, mapas ou combinações destes que constituem o cadastro devem ser examinados para que seja assegurado que estejam livres de defeitos e, caso estejam desatualizados, deve-se considerar a possibilidade de proceder sua atualização. Em qualquer caso, é importante conhecer a origem do cadastro utilizado. A especificação e composição do cadastro são partes importantes das fases de planejamento do processo de produção de uma pesquisa e a atualidade e qualidade do cadastro têm grande influência nos resultados da pesquisa. Um bom *plano amostral* depende criticamente da qualidade do cadastro disponível e de seu conhecimento pelos responsáveis pela elaboração do plano.

Na etapa de planejamento da amostra, questões tais como a definição do tamanho da amostra, a maneira de selecionar as unidades que irão compor a amostra e a especificação de estimadores para os parâmetros populacionais de interesse, bem como para as respectivas margens de erro, são problemas técnicos que devem merecer a mais cuidadosa atenção. De fato, as ideias e técnicas para resolver estas questões formam o principal conteúdo deste livro.

Os métodos a serem usados para a coleta das informações dependem de muitos fatores, e devem ser decididos tendo em mente os custos envolvidos, os tipos de unidades informantes, a precisão desejada das medidas e observações, bem como as condições particulares para execução da pesquisa em questão. Tais métodos são geralmente especificados durante a fase de planejamento da pesquisa, pois os métodos de coleta de dados influenciarão fortemente na elaboração do questionário da pesquisa e também podem afetar as opções disponíveis para o planejamento da amostra.

Há diferentes métodos de coleta, que podem ou não ser apoiados por computador e podem ou não necessitar da participação de entrevistadores ou observadores. Os principais métodos de coleta são: *Computer-Assisted Personal Interviewing* - CAPI, *Computer-Assisted Self Interviewing* - CASI, *Computer-Assisted Telephone Interviewing* - CATI, *Mail Assisted Self Interviewing* - MASI, *Paper-and-Pencil Interviewing* - PAPI e métodos de observação direta, em que não ocorre uma entrevista. Um exemplo de aplicação deste último método ocorre na coleta de dados de preços em pontos de venda, onde o responsável pela coleta de dados faz isso diretamente e não requer contato com uma unidade informante para a obtenção das informações de interesse. Na *Pesquisa Nacional de Saúde do Escolar* - PeNSE, realizada pelo IBGE, é utilizado o método CASI, onde cada um dos estudantes selecionados para

a amostra recebe um *tablet* com o questionário e o responde sem interferência de um entrevistador. Em tempos mais recentes, também começaram a tornar-se disponíveis opções tais como *web-scraping*, que consiste na utilização de métodos automatizados de acesso a páginas da internet e de extração de dados ou informações dessas páginas, sem intervenção humana no processo.

1.4.3 Elaboração dos instrumentos de coleta e dos sistemas

Esta etapa envolve os subprocessos que definem os instrumentos de coleta, os sistemas de processamento, de disseminação e de produção da pesquisa.

Os instrumentos de coleta (questionários e outros formulários) são parte importante de uma pesquisa por amostragem. Após decidir que dados devem ser coletados, a questão de como formular e apresentar as perguntas exige grande habilidade e prática. As perguntas devem ser claras, inambíguas e diretas. Perguntas vagas tendem a ter respostas vagas. A ordenação das perguntas deve ser estudada com cuidado. Um pré-teste dos instrumentos de coleta é sempre uma ajuda efetiva na preparação de um bom material.

Um protocolo de esforço de coleta deve ser especificado. Este protocolo deve incluir, por exemplo, instruções sobre o número mínimo de tentativas de contato que devem ser feitas para cada unidade da amostra selecionada, antes de dar uma unidade como perdida. Os procedimentos para lidar com recusas ou perdas de unidades da amostra selecionada também devem ser definidos previamente. Os motivos das perdas devem sempre ser registrados e o tratamento para as perdas e para a não resposta devem ser especificados.

Nesta etapa estão ainda incluídos a elaboração e os testes dos sistemas de apuração e de disseminação da pesquisa.

1.4.4 Coleta dos dados

Esta etapa inclui a implementação da pesquisa e a coleta dos dados, na qual se destacam as seguintes atividades: a preparação do cadastro, a seleção da amostra, o recrutamento e treinamento de equipes de coleta e supervisão de campo e a realização e o acompanhamento da coleta.

Para os propósitos da seleção da amostra, a população deve poder ser subdividida no que se pode chamar de *unidades de amostragem*. É importante que tal divisão seja feita de forma que nenhuma unidade da população pertença a mais de uma unidade de amostragem ou fique de fora do conjunto de todas as unidades de amostragem. Um exemplo pode ser dado com as populações humanas, que podem ser vistas como formadas por agregações tais como setores censitários e domicílios ou por pessoas.

A seleção da amostra se dá de acordo com o plano (desenho) amostral especificado a partir do cadastro a ser usado e varia de uma amostra aleatória simples (ou sistemática), podendo ser estratificada, ao uso de amostragem com probabilidades proporcionais ao tamanho. Outras alternativas de planos amostrais contemplam a amostragem de conglomerados(agrupamentos de unidades populacionais) ou em múltiplos estágios, motivada pela necessidade de eficiência prática e econômica. Os principais métodos de amostragem são abordados nos capítulos subsequentes.

O processo de coleta de dados envolve as estratégias de contato e de monitoramento. O sucesso de qualquer pesquisa que adote o método de entrevista direta depende fortemente da capacidade dos entrevistadores de obter as respostas desejadas. Por isso, sua seleção e treinamento é muito importante. Instruções detalhadas devem ser fornecidas no treinamento sobre os métodos de mensuração e coleta. A observação por supervisores durante os trabalhos de entrevista é fundamental para manter os padrões e para estudar a obediência às regras e o tato dos entrevistadores ao fazer as entrevistas. É fundamental que todos os entrevistadores entendam os conceitos da pesquisa de maneira correta e uniforme.

1.4.5 Processamento da pesquisa

Esta etapa envolve os subprocessos que definem a integração/organização dos dados, classificação e codificação, crítica e tratamentos dos dados, derivação de variáveis, cálculo dos pesos ou fatores de expansão da amostra, cálculo dos resultados agregados (estimação das quantidades de interesse) e preparação dos arquivos de dados.

Ao final da fase de coleta de dados, as informações estão prontas para entrar na fase do processamento, quando os registros de dados são verificados, criticados e preparados para a análise. A maneira como as etapas a seguir são encadeadas e executadas depende de como a coleta de dados é organizada e do(s) modo(s) usado(s) para coletar as informações.

A codificação consiste em atribuir código numérico a respostas obtidas inicialmente em forma de texto, por meio de uma classificação predeterminada. É o caso, por exemplo, da ocupação da pessoa cuja descrição dada pelo informante é transformada num código que é estruturado por uma classificação. É mais fácil o informante responder fornecendo uma descrição registrada em texto (pergunta aberta) e, em seguida, interpretar essa resposta para alocação de um código ou classe com base na classificação adotada para o tema na pesquisa. Atualmente, a maior parte do trabalho de codificação é feita mediante a combinação de codificação automática, quando é possível associar um único código às respostas textuais, com um sistema de codificação assistida por computador (acionado por operadores), para aqueles casos nos quais nenhum ou mais de um código foi encontrado pelo sistema automático para uma dada resposta.

A crítica e tratamento dos dados coletados é uma etapa indispensável para permitir a eliminação de erros grosseiros na massa de dados coletados, os quais podem distorcer significativamente os resultados da pesquisa. É preciso ter formas de detectar inconsistências e definir o tratamento para a correção dos dados individuais. Um bom texto de referência sobre o tema é o livro de Waal et al. (2011).

Uma situação que quase sempre ocorre em pesquisas estatísticas é que os dados coletados são incompletos, em função da ocorrência de valores ausentes, seja por não resposta ou por terem sido descartados por inconsistências detectadas no processo de crítica. A obtenção de um conjunto de dados com registros completos antes da etapa de estimação se dá através de imputação, substituindo os valores ausentes ou descartados em registros incompletos por valores estimados com base nos dados disponíveis. Vários métodos estão disponíveis para imputar os valores ausentes em um conjunto de dados, dentre os quais podem ser citados: imputação dedutiva, imputação baseada em modelo (incluindo imputação por média, razão e regressão) e imputação por registro doador. Ver, a respeito, os excelentes livros de Waal et al. (2011), Little e Rubin (2002), Rubin (1987) e Schafer (1997).

Os métodos de estimação são usados para generalizar a informação recolhida de uma amostra para a população da qual foi extraída. A forma de selecionar a amostra determina como serão produzidas tais estimativas da população. De fato, o plano amostral determina os chamados *pesos amostrais* (fatores de expansão) *básicos* que serão usados para produzir estimativas. O *peso amostral* de uma unidade observada indica o número de unidades da população que são representadas por esta unidade da amostra. Um *peso amostral básico* é calculado como o inverso da probabilidade de incluir a unidade na amostra.

Os procedimentos de estimação das quantidades de interesse envolvem especificar: o cálculo dos pesos ou fatores de expansão; os estimadores para as quantidades de interesse; e o cálculo das medidas de precisão das estimativas.

1.4.6 Análise dos dados da pesquisa

Esta etapa envolve a análise e interpretação dos dados. Usualmente, a apresentação de resultados é acompanhada de um exercício analítico inicial dos dados produzidos. A análise de dados é o processo pelo qual se dá ordem, estrutura, interpretação e significado aos dados.

A análise permite diferentes abordagens e requer cuidado na interpretação, destacando-se que devem ser observadas as seguintes características: que seja apresentada de maneira simples, objetiva e compreensível, com significado claro; e que busque assegurar que as informações produzidas e analisadas sejam úteis para satisfazer aos objetivos enunciados da pesquisa. Uma boa forma de avaliar isso é verificar se os comentários da análise fornecem respostas às principais perguntas formuladas para justificar a realização da pesquisa.

A informação nunca é completamente útil se não for acompanhada de análise. Neste caso, o foco das atenções deve ser dado sobre as questões de coerência e interpretabilidade e em dar significado aos resultados em sua apresentação. A interpretabilidade das estatísticas também faz parte da noção de qualidade dos dados produzidos. Estatísticas definidas de maneira hermética ou complexa, cuja interpretação seja difícil, raramente conseguem estabelecer uma percepção de boa qualidade.

Esta etapa também envolve a garantia de que os dados (e metadados) a serem disseminados não violem a confidencialidade de informações individualizadas, prevenindo a divulgação de informações que revelem direta ou indiretamente a identidade do informante (indivíduo, empresa ou instituição) com a associação de dados confidenciais. Devem ser usadas técnicas para limitação da revelação da identidade nas diversas formas de acesso aos dados.

1.4.7 Disseminação dos resultados da pesquisa

Esta etapa envolve a produção dos resultados, a aplicação de métodos para assegurar a confidencialidade das informações individualizadas, a promoção dos produtos de disseminação e o atendimento aos usuários. Também envolve a preparação do material de divulgação - tabelas, textos, apresentações, publicações, arquivos de microdados e sua documentação (metadados), elaboração de releases e carga em sítios da internet. É das etapas menos discutidas do processo de pesquisa nos textos sobre métodos, mas isto não significa que é pouco importante. Na verdade, pode-se afirmar que sem adequada disseminação se perde muito do valor de uma pesquisa, já que os seus resultados ficarão subaproveitados.

1.4.8 Arquivamento das informações da pesquisa

Esta etapa envolve os subprocessos que definem as regras de arquivamento, o gerenciamento do repositório, a preservação dos dados, metadados e paradados e o plano de descarte de material.

A busca e recuperação da informação apontam para a criação de metadados (que descreve os dados de forma estruturada) e a preservação das informações, permitindo que sejam acessíveis e com a autenticidade resguardada. Neste sentido as regras de arquivamento e o gerenciamento do repositório têm um papel relevante na estruturação das informações de forma a localizá-las e torná-las de fácil recuperação e uso.

1.4.9 Avaliação da pesquisa

Esta etapa encerra a execução de uma pesquisa e o produto final deve ser um conjunto de tabulações e relatórios de avaliação da pesquisa e de seus resultados. A avaliação e o monitoramento da qualidade das estimativas deve estar presente desde a concepção da pesquisa, a operacionalização dos processos, até a elaboração e disseminação do produto final, visando a compreensão das informações pelos usuários.

Cabe registrar a ideia de que qualidade é 'multidimensional' e que há vários aspectos da produção dos dados que se deve considerar para avaliar a qualidade dos resultados.

Começamos por distinguir dois níveis onde a discussão de qualidade é importante. Num primeiro nível, trata-se de qualidade do processo de produção das estatísticas, que está relacionada ao desempenho da organização produtora, à sua capacidade de adotar métodos e processos de trabalho eficientes e seguros, bem como de responder às demandas que lhe são apresentadas. Um segundo nível se refere à qualidade dos dados e resultados que fazem parte da produção das estatísticas. Neste momento, a qualidade que importa é a de cada resultado (produto) frente a um conjunto de usos previstos ou antecipados que o resultado terá.

Cabe mencionar algumas das dimensões da qualidade associadas à qualidade do produto: relevância, exatidão e confiabilidade, oportunidade e pontualidade, facilidade de acesso e uso, interpretabilidade, coerência, dentre outras.

A questão da relevância é colocada quando se avalia se o produto (resultado, informação estatística, estimativa, dado) satisfaz algum uso (ou conjunto de usos) declarado legítimo e importante pelos diversos atores interessados na produção das estatísticas.

A questão da exatidão e confiabilidade coloca em evidência a qualidade das estatísticas de um ponto de vista mais familiar para quem faz e discute Estatística. Exatidão de uma estatística tem a ver com a ausência de viés (vício) na estimação da medida de interesse. Confiabilidade (precisão) tem a ver com o erro máximo provável cometido ao estimar a quantidade de interesse usando os dados disponíveis.

A noção de qualidade como exatidão ou confiabilidade pode ser quantificada ou medida, para um produto ou resultado qualquer. Por esse motivo, muitas vezes este aspecto da qualidade costuma receber uma atenção maior na definição ou interpretação de qualidade que outros aspectos, o que nem sempre é justificável. Por exemplo, em pesquisas por amostragem bem planejadas e executadas, com amostras grandes, estimativas de muitos parâmetros obtidas para o conjunto da população estudada terão exatidão assegurada pelo planejamento e alta confiabilidade e, nesses casos, outras dimensões da qualidade talvez devessem assumir papel de maior destaque na análise.

Também é de grande importância a questão da oportunidade e pontualidade das informações. Estatísticas divulgadas com atraso geralmente suscitam desconfiança. Estatísticas referentes a períodos muito distantes no tempo podem ter utilidade limitada ou até levar a erro, pois a realidade a que se referiam pode já ter mudado substancialmente. Desta forma, é importante para as agências produtoras de estatísticas oficiais trabalhar com calendários de divulgação de resultados previamente divulgados e cumprir esses calendários, ao mesmo tempo em que se esforcem para disseminar

os resultados de cada pesquisa o mais cedo possível após a coleta das informações. Estes dois objetivos devem ser perseguidos, entretanto, tendo como contraponto a ideia de que revelar de forma apressada estatísticas sujeitas a grandes revisões e correções posteriores não conduz a uma percepção de qualidade no trabalho da agência produtora de estatísticas.

Embora sem esquecer os demais aspectos de uma pesquisa por amostragem, enfatizamos neste livro as técnicas e métodos para: seleção da amostra; estimação dos parâmetros desejados; e avaliação dos erros de amostragem. O tratamento de erros ditos *não amostrais* não é objeto de atenção aqui. Aos leitores interessados, recomendamos a leitura do excelente livro de Biemer e Lyberg (2003).

1.5 Cadastros

O *Cadastro* ou *Sistema de Referência* da pesquisa é o conjunto de meios e instrumentos que fornece o acesso à população de pesquisa, constituindo uma lista identificadora das unidades que formam a população ou de grupos destas unidades e contendo informações auxiliares úteis para planejar e selecionar a amostra, monitorar a coleta de informações e empregar na estimação dos parâmetros.

A especificação do cadastro é uma das partes mais importantes da fase de planejamento do processo de produção da pesquisa. Nesta fase devem ser definidas as *unidades de referência* e o escopo da *população de pesquisa*, as *unidades informantes* e o processo que será usado para o acesso a estas unidades. O período de referência e as condições de seleção e atribuição das unidades do cadastro também precisam ser especificados. O cadastro deve conter os atributos das unidades demandadas pelas fases de amostragem, coleta e processamento dos dados.

Para a seleção da amostra, o cadastro fornece, além do contato com as unidades informantes, informações para estratificação da população. Além disso, ajuda a controlar e monitorar a fase de coleta de dados, a registrar e validar as respostas, a estimular a colaboração dos respondentes e a avaliar as eventuais não respostas. O cadastro também fornece informações para as fases de ponderação da amostra e análise dos resultados.

Assim, o cadastro utilizado afeta diretamente a definição da população de pesquisa, o método de coleta de dados, o método de seleção da amostra, a qualidade dos resultados e o custo da pesquisa.

Há vários tipos de cadastros: *de unidades individuais*, constituído por uma lista física ou conceitual das unidades de referência individuais que formam a população; *de áreas*, com correspondentes mapas e descrições de áreas geográficas, das quais são selecionadas áreas para a amostra, cujas unidades de referência correspondentes são enumeradas e eventualmente amostradas para a pesquisa; *de conglomerados*, constituídos por listas de grupos de unidades individuais tais como escolas (agrupando alunos), hospitais (agrupando pacientes), etc.; e *cadastros múltiplos*, que combinam dois ou mais cadastros, de mesmo tipo ou não.

Uma pesquisa pode ter mais de um cadastro. Em pesquisas com seleção da amostra em múltiplos estágios, geralmente há um cadastro para cada estágio. Por exemplo, na Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua, o cadastro do primeiro estágio é constituído pelos setores, enquanto o cadastro de segundo estágio é formado pelos domicílios enumerados em cada setor selecionado no primeiro estágio.

Os registros administrativos também são fontes importantes, pois registram as unidades estatísticas e uma ampla gama de suas variáveis é usada para a criação do cadastro. Sua atualidade e qualidade têm grande influência na qualidade da pesquisa e nos produtos estatísticos.

Um bom cadastro deve conter informação suficiente sobre cada unidade da população para identificá-la com certeza (identificação) e para permitir localizá-la (localização); ser completo e sem redundâncias (duplicatas), preciso e atual (qualidade); estar disponível em um lugar central, com acesso fácil e rápido (disponibilidade); estar arranjado numa forma adequada à amostragem; e conter informação auxiliar sobre cada unidade que permita elaborar um planejamento amostral e estratégias de estimação eficientes.

1.5.1 Cadastros importantes no IBGE

O *Cadastro Central de Empresas - CEMPRE* é formado por empresas e outras organizações e suas respectivas unidades locais formalmente constituídas, registradas no CNPJ - Cadastro Nacional da Pessoa Jurídica. Sua atualização ocorre anualmente, a partir das pesquisas econômicas anuais do IBGE, nas áreas de Construção, Indústria, Comércio e Serviços, e de registros administrativos tais como a Relação Anual de Informações Sociais - RAIS e o Cadastro Geral de Empregados e Desempregados - CAGED, ambos do Ministério do Trabalho.

As informações disponíveis referem-se às empresas e às unidades locais que no ano de referência estavam ativas no Cadastro. Estão disponíveis as variáveis identificadoras de empresas e de suas unidades locais, o pessoal ocupado total, o pessoal assalariado, os salários e outras remunerações pagas e o salário médio mensal, além da Classificação Nacional de Atividades Econômicas - CNAE e da localização de cada unidade.

A cada ano, é extraído do CEMPRE um *Cadastro Básico de Seleção* usado na seleção das amostras das pesquisas econômicas anuais, tais como a Pesquisa Anual da Indústria da Construção - PAIC, a Pesquisa Industrial Anual - PIA, a Pesquisa Anual do Comércio - PAC e a Pesquisa Anual de Serviços - PAS. Além destas pesquisas econômicas anuais, o CEMPRE também serve de base para a extração de amostras das pesquisas conjunturais do IBGE, tais como a Pesquisa Industrial Mensal de Produção Física - PIM-PF, a Pesquisa Mensal do Comércio - PMC e a Pesquisa Mensal dos Serviços - PMS.

Algumas outras pesquisas feitas por outras organizações também se valem do CEMPRE para seleção de suas amostras. Este é o caso da Pesquisa TIC-Empresas do NIC.br - ver detalhes em https://cetic.br/pesquisa/empresas/publicacoes.

A *Base Operacional Geográfica - BOG* é um cadastro de áreas que tem como suas menores unidades os setores censitários e compreende uma hierarquia de unidades geoestatísticas, aqui listadas da menor para a maior: setores censitários, subdistritos, distritos, municípios, unidades da federação, macrorregiões. Foi construída e é mantida para dar organização e sustentação espacial às atividades de planejamento, coleta, apuração e divulgação dos resultados do Censo Demográfico e do Censo Agropecuário, bem como para o planejamento e execução das pesquisas domiciliares (Pesquisa Nacional por Amostra de Domicílios Contínua, Pesquisa de Orçamentos Familiares, etc.) - ver IBGE (2016).

O *setor censitário* é a unidade territorial de controle cadastral da coleta, constituída por áreas contíguas, respeitando os limites da divisão político-administrativa, do quadro urbano e rural legal e de outras estruturas territoriais de interesse, além dos parâmetros de dimensão mais adequados à operação de coleta do Censo Demográfico. Ver, a respeito, IBGE (2016).

A codificação (numeração única de cada setor censitário), a definição do tamanho, a classificação segundo a situação (urbana ou rural) e o tipo (comum ou não especial, aglomerado subnormal, quartel ou base militar, etc.), a genealogia e a descrição dos limites dos setores estão registrados na BOG, que através destas informações fornece os instrumentos essenciais para o controle das operações de coleta do Censo e de pesquisas domiciliares. Para detalhes sobre os tipos de setores censitários, ver IBGE (2019), página 169.

A malha digital de setores censitários do Brasil é um conjunto de arquivos contendo os polígonos definidores de estados, municípios, distritos, subdistritos, bairros e setores censitários. Está disponível juntamente com os dados agregados do Censo Demográfico 2010, por setor censitário.

A Figura 1.2 apresenta uma ilustração da subdivisão em setores censitários do IBGE para Copacabana no município do Rio de Janeiro, conforme a malha setorial vigente para o Censo Demográfico 2010.

O *Cadastro Nacional de Endereços para Fins Estatísticos - CNEFE* é uma lista com cerca de 78 milhões de endereços urbanos e rurais, associados às unidades (domicílios e unidades não residenciais) registradas pelos recenseadores durante a coleta das informações do Censo Demográfico 2010, e aos setores censitários. Foi compilado para apoiar a realização das pesquisas domiciliares do IBGE.

As Figuras 1.3 e 1.4 ilustram informações disponíveis no CNEFE para um determinado setor censitário. Ver detalhes no link https://censo2010.ibge.gov.br/cnefe/Exibe_Tabela.html?ag=330455705100417.

A seguir, os cadastros associados à BOG do IBGE que estão disponíveis para uso público são listados juntamente com seus endereços de acesso via internet.

Arquivo Agregado de Setores (Censo 2010)

https://www.ibge.gov.br/estatisticas-novoportal/downloads-estatisticas.html »» Censos »» Censo_Demográfico_2010 »» Resultados_do_Universo »» Agregados_por_Setores_Censitarios

Malha Digital de Setores Censitários (Censo 2010)

https://www.ibge.gov.br/estatisticas-novoportal/downloads-estatisticas.html »» Geociências »» Organização_do_território malhas_territoriais »» malhas_de_setores_censitarios_divisoes_intramunicipais »» censo_2010 »» setores_censitarios_shp

Cadastro Nacional de Endereços para Fins Estatísticos - CNEFE

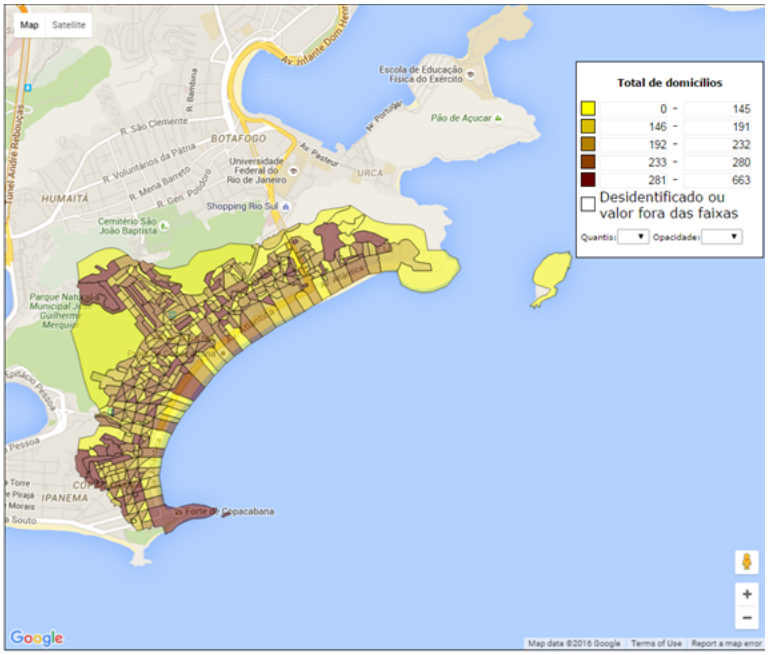
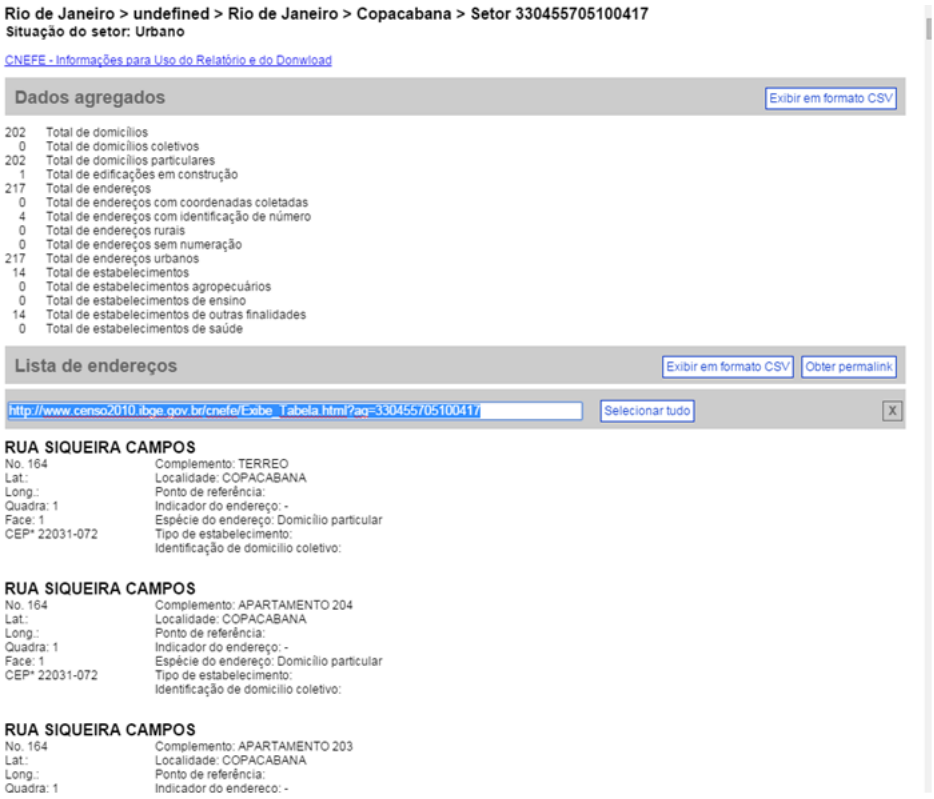


Figura 1.2: Cadastro de setores do IBGE - Copacabana - RJ



Rio de Janeiro > undefined > Rio de Janeiro > Copacabana > Setor 330455705100417
Situação do setor: Urbano

CNEFE - Informações para Uso do Relatório e do Donwload

Dados agregados — Exibir em formato CSV

202	Total de domicílios
0	Total de domicílios coletivos
202	Total de domicílios particulares
1	Total de edificações em construção
217	Total de endereços
0	Total de endereços com coordenadas coletadas
4	Total de endereços com identificação de número
0	Total de endereços rurais
0	Total de endereços sem numeração
217	Total de endereços urbanos
14	Total de estabelecimentos
0	Total de estabelecimentos agropecuários
0	Total de estabelecimentos de ensino
14	Total de estabelecimentos de outras finalidades
0	Total de estabelecimentos de saúde

Lista de endereços — Exibir em formato CSV | Obter permalink

http://www.censo2010.ibge.gov.br/cnefe/Exibe_Tabela.html?aq=330455705100417 — Selecionar tudo — X

RUA SIQUEIRA CAMPOS
No. 164 — Complemento: TERREO
Lat.: — Localidade: COPACABANA
Long.: — Ponto de referência:
Quadra: 1 — Indicador do endereço: -
Face: 1 — Espécie do endereço: Domicílio particular
CEP* 22031-072 — Tipo de estabelecimento:
Identificação de domicílio coletivo:

RUA SIQUEIRA CAMPOS
No. 164 — Complemento: APARTAMENTO 204
Lat.: — Localidade: COPACABANA
Long.: — Ponto de referência:
Quadra: 1 — Indicador do endereço: -
Face: 1 — Espécie do endereço: Domicílio particular
CEP* 22031-072 — Tipo de estabelecimento:
Identificação de domicílio coletivo:

RUA SIQUEIRA CAMPOS
No. 164 — Complemento: APARTAMENTO 203
Lat.: — Localidade: COPACABANA
Long.: — Ponto de referência:
Quadra: 1 — Indicador do endereco: -

Figura 1.3: Informações do CNEFE para um setor de Copacabana - RJ

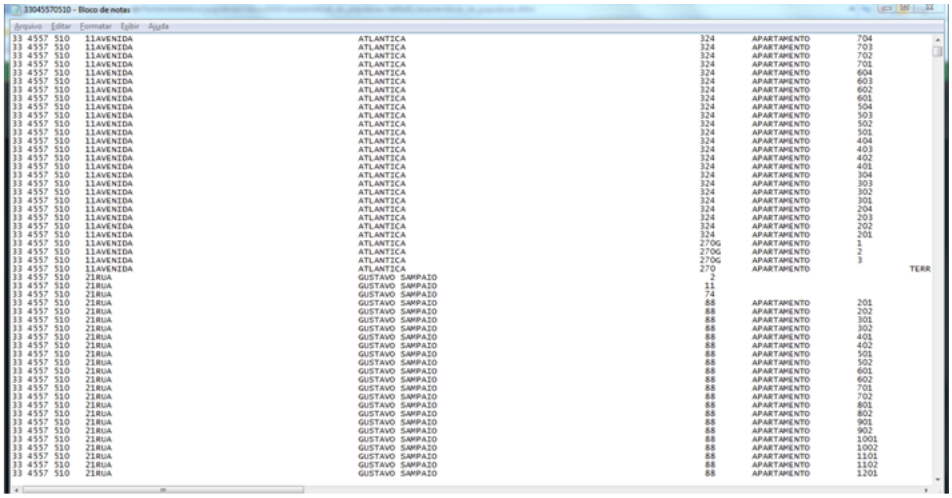


Figura 1.4: Ilustração de dados do CNEFE

https://www.ibge.gov.br/estatisticas-novoportal/downloads-estatisticas.html »» Censos »» Censo_Demográfico_2010 »» Cadastro_Nacional_de_Enderecos_Fins_Estatisticos

1.5.2 Defeitos de cadastros

Para determinadas populações, a lista de elementos pode estar disponível, mas pode apresentar alguns defeitos: falta de unidades (omissão ou falha de cobertura), presença de unidades estranhas à população alvo (unidades inelegíveis ou fora do escopo da pesquisa), duplicação de unidades, informações desatualizadas, informações faltando ou incorretas (por exemplo, unidades com os dados de contato incorreto ou incompleto).

As inadequações do cadastro podem levar a problemas operacionais na coleta e no processamento dos dados da pesquisa, aumento dos erros alheios à amostragem e interpretações enganosas dos resultados da pesquisa.

Diante das inadequações do cadastro, dentre as soluções possíveis destacam-se: o descarte do cadastro e a criação ou uso de outro cadastro; o ajuste/correção do cadastro mediante atualização ou ligação com outros; o uso do cadastro existente e adoção de precauções contra seus defeitos; e uso de cadastros múltiplos. Um tratamento detalhado das providências disponíveis está fora do escopo deste livro, mas como já indicado, recomendamos dar grande atenção à questão da seleção e preparação do cadastro que vai apoiar as atividades de pesquisa por amostragem, dado seu papel central na obtenção de resultados de boa qualidade.

1.5.3 Regras de associação das unidades da população à unidade cadastral

O cadastro deve ser estruturado de tal forma que seja possível determinar como as unidades listadas no cadastro estão associadas às *unidades de referência* na população de pesquisa a ser amostrada. Na sequência, apresentamos as várias formas de associação entre *unidades do cadastro* e *unidades de referência* da população de pesquisa. Cada uma das situações leva à adoção de métodos de amostragem que considerem as diferentes formas de associação entre cadastro e população.

Regra de associação um para um

Nesta situação cada unidade C_i do cadastro corresponde a uma e somente uma unidade de referência U_i da população de pesquisa, conforme ilustração na Figura 1.5.

Neste caso, a seleção da amostra de unidades elementares pode ser feita diretamente do cadastro. Os planos amostrais podem selecionar diretamente *unidades de referência* elementares e não há conglomeração. Portanto, a *unidade de referência* é também a *unidade de amostragem*. A seleção da amostra fica bem simplificada, porém a manutenção do cadastro costuma ser mais cara quando comparada a outros tipos de situações e a cobertura é mais díficil de ser mantida.

Exemplo 1.7. Cadastro com regra de associação um para um

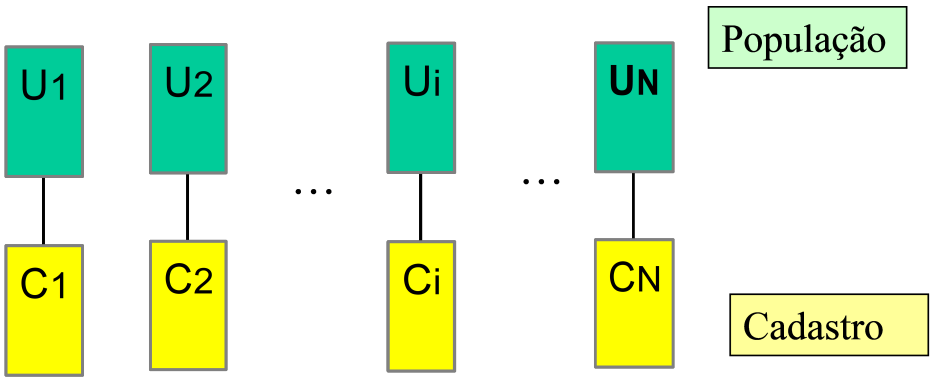


Figura 1.5: Ilustração da regra de associação um para um

Um exemplo clássico de cadastro com esse tipo de associação é o *Cadastro Básico de Seleção* extraído do CEMPRE e usado pelo IBGE para seleção das amostras de suas pesquisas econômicas estruturais (PAIC, PIA, PAC e PAS). Nesse tipo de cadastro, a unidade elementar é uma empresa, que corresponde também à unidade de referência das pesquisas citadas. Os planos amostrais adotados nestas pesquisas tomam a empresa como unidade de amostragem.

Regra de associação um para vários

Nesta situação, cada unidade de referência da população de pesquisa corresponde a uma ou mais unidades do cadastro, conforme ilustração na Figura 1.6. Consideramos aqui apenas os casos em que cada unidade elementar no cadastro tenha vínculo com no máximo uma unidade elementar na população.

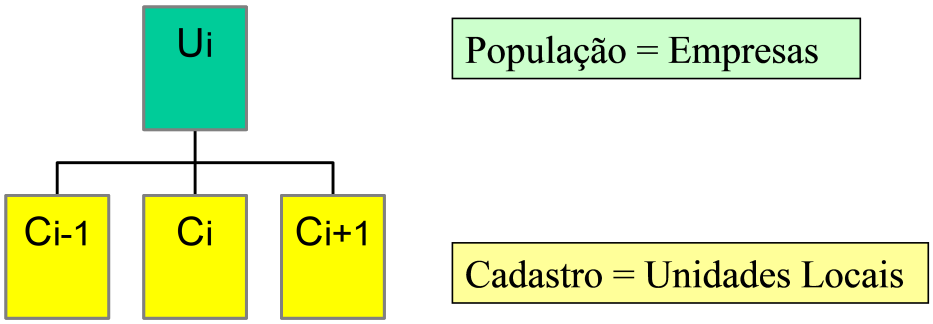


Figura 1.6: Ilustração da regra de associação um para vários

Neste caso, a *unidade de amostragem* é a unidade elementar do cadastro e a seleção da amostra é feita diretamente do cadastro. As probabilidades de seleção das unidades de referência da população de pesquisa variam com o número de unidades elementares correspondentes no cadastro. A manutenção do cadastro e dos vínculos é geralmente trabalhosa e o conhecimento exaustivo e preciso dos vínculos é essencial. Note que também não se aplica aqui a ideia de amostragem conglomerada, já que as unidades de referência da população são incluídas (ou não) na amostra uma a uma, dependendo do sorteio de unidades do cadastro com que estão vinculadas.

Exemplo 1.8. Cadastro com regra de associação um para vários

Outro bom exemplo deste tipo de situação seria um cadastro de veículos (automóveis) particulares de pessoas físicas, onde estão registrados os veículos e seus proprietários podem ser identificados através do Cadastro de Pessoas Físicas - CPF. Imagine que há interesse em selecionar uma amostra de proprietários de veículos, que deve ser extraída usando o cadastro de veículos. Estamos diante de uma situação em que a cada proprietário de veículo que pertence à população de pesquisa estão associados um ou mais veículos. Então fica configurada a situação em que o veículo é a unidade de amostragem, que não coincide com a unidade de referência de interesse da pesquisa, que é o proprietário de veículo. Caso seja feita amostragem simples (com igual probabilidade) dos veículos, os proprietários terão probabilidades de seleção para a amostra proporcionais ao número de veículos que tenham registrados no cadastro de veículos.

Regra de associação de vários para um

Nesta situação, uma ou mais unidades de referência da população de pesquisa são vinculadas a cada unidade elementar do cadastro, conforme ilustrações nas Figuras 1.7 e 1.8. A unidade de amostragem é um *conglomerado* de unidades da população. Neste caso, se adotam os planos amostrais conglomerados, onde a seleção é de uma amostra de unidades conglomeradas do cadastro. Os vínculos são geralmente conhecidos só para os conglomerados da amostra. O cadastro é mais barato de construir e manter, porém a amostragem é na maioria das vezes menos eficiente do que se poderia fazer tendo cadastros do tipo um para um.

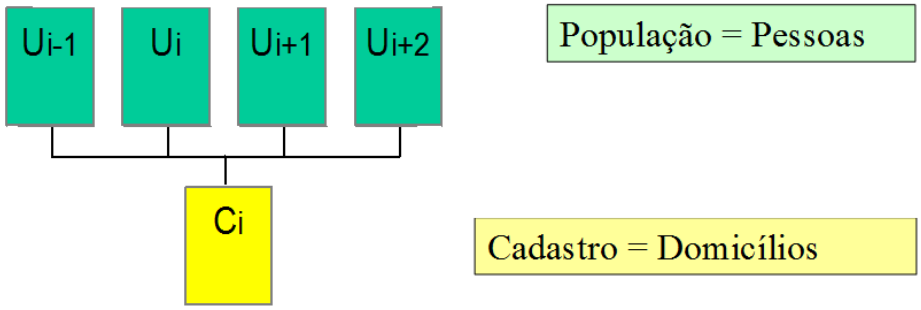


Figura 1.7: Ilustração da regra de associação vários para um

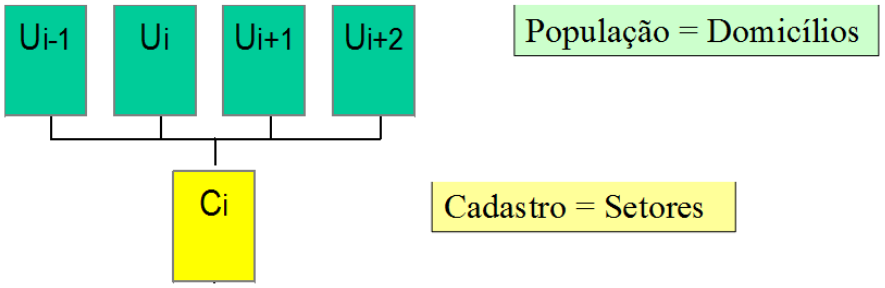


Figura 1.8: Ilustração da regra de associação vários para um

Regras de Associação - Hierarquia com vários níveis

Neste caso se adotam os planos de amostragem conglomerada em múltiplos estágios (ou etapas). Envolve uma hierarquia de diferentes tipos de unidades (unidades primárias de amostragem, unidades secundárias, terciárias, etc.). É necessário um bom cadastro em cada estágio para a seleção das unidades do estágio seguinte. Os cadastros para as unidades dos primeiros estágios são geralmente mais estáveis e fáceis de construir e manter que aqueles para os estágios subsequentes. A Figura 1.9 apresenta uma ilustração dessa situação.

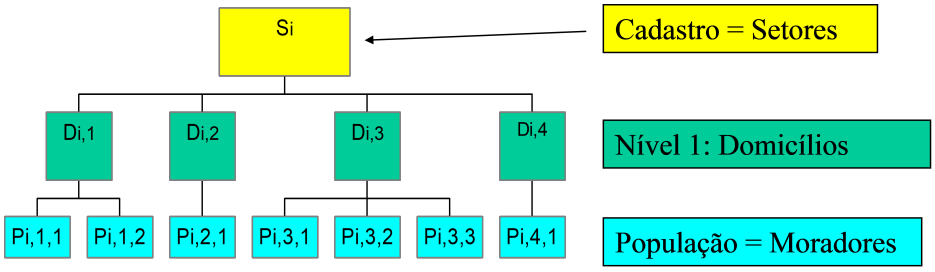


Figura 1.9: Ilustração de regras de associação - hierarquia para vários níveis

1.6 Exercícios

Exercício 1.1. Toda pesquisa deve procurar registrar com clareza seus conceitos básicos. Dessa clareza depende a capacidade dos usuários de interpretar os resultados e avaliar sua qualidade e adequação. Para perceber a importância dessa questão, nada melhor do que praticar e se colocar "na pele" de um usuário que procura identificar, na documentação disponível de uma dada pesquisa, seus principais conceitos e definições.
Visite a página do CETIC.br na internet (https://cetic.br/) e identifique a publicação de resultados da Pesquisa TIC Domicílios 2017. Nesta publicação, localize a seção do 'relatório metodológico' e usando somente essa documentação, procure responder às perguntas abaixo.

a) Qual é o *objetivo* da pesquisa?

b) Qual é a *população alvo*? Qual é a *população de pesquisa*?

c) Qual é o principal *cadastro* utilizado?

d) Quais são as *unidades de referência*?

e) Quais são as *unidades informantes*?

f) Quais são as *unidades de análise*?

g) Quais são as *unidades de amostragem*?

h) Qual é o *tipo* de pesquisa? *Censo*, *amostra*, compilação de *registro administrativo*?

i) Como a pesquisa se desenvolve no *tempo*? Uma só ocasião? Pesquisa Repetida? Se a pesquisa é repetida, qual é a periodicidade com que isso ocorre?

j) Como são *coletados* os dados? Entrevistas, correio, observação, etc.?

k) Para que tipos de *quantidades resumo* são publicadas estimativas? Totais, médias, proporções, razões, taxas, índices? Exemplifique alguns que encontrar.

l) Foi possível responder todas as perguntas acima só com a documentação da pesquisa disponível na internet? Comente sobre suas dificuldades. O que ficou faltando?

Exercício 1.2. Um pesquisador está planejando a realização de uma pesquisa por amostragem junto a estudantes do ensino fundamental regular - EFR matriculados em *escolas públicas municipais* no município do Rio de Janeiro. A coleta dos dados estava prevista para ser realizada entre *setembro e outubro de 2019*. Ele começou a investigar opções de cadastros que poderia utilizar para realizar a seleção da amostra. De imediato, localizou os seguintes cadastros:

a) *Cadastro* derivado do *Censo Escolar do MEC*, referente ao ano calendário *2018*, que pode ser acessado no endereço: http://portal.inep.gov.br/web/guest/microdados.

b) *Lista* de todas as *escolas públicas municipais* que oferecem turmas do ensino fundamental regular fornecida pela Prefeitura municipal, atualizada até Agosto de 2019, contendo nome e endereço da escola, nome e dados de contato do diretor da escola, e séries que a escola oferece no ano de 2019.

Tratar das seguintes questões:

1) Para cada um dos cadastros disponíveis para o pesquisador, indique o tipo de associação entre as unidades do cadastro e as unidades de referência da pesquisa (estudantes do ensino fundamental regular matriculados em escolas públicas municipais em setembro de 2019).

2) Discuta as vantagens e desvantagens de utilizar cada um dos cadastros disponíveis para o pesquisador.

3) Caso o pesquisador opte pelo uso do cadastro listado no item a, que cuidados deveria adotar para mitigar ou eliminar possíveis erros de cobertura?

4) Caso o pesquisador opte pelo uso do cadastro listado no item b, que procedimento(s) precisa adotar para assegurar a cobertura da população de interesse?

5) Se você fosse o pesquisador, como usaria um ou mais dos cadastros citados?

Capítulo 2

Visão Geral da Amostragem e Estimação

2.1 Definições e notação para população de pesquisa e parâmetros selecionados

Nesta seção introduzimos algumas definições e notação necessárias para a apresentação da teoria da amostragem ao longo do texto.

Chama-se *população* (daqui por diante, esta é a designação da *população de pesquisa* que será objeto do levantamento de dados) qualquer conjunto contendo um número finito N de unidades, delimitada por compartilharem algumas características em comum. As unidades deste conjunto são denominadas *unidades da população*. Representamos nosso cadastro dessa população por um conjunto de N rótulos distintos denotado $U = \{1,\ 2,\ ...,\ i,\ ...,\ N\}$, sendo N o *tamanho da população* e i o rótulo para uma unidade genérica da população. Cada unidade da população fica devidamente identificada por seu rótulo no conjunto U.

São exemplos comuns de populações sobre as quais se realizam pesquisas: domicílios e moradores de certa localidade; indústrias instaladas num certo país; fazendas situadas num certo estado; alunos matriculados na 3a série do ensino médio da rede escolar estadual em 2017.

No capítulo anterior já enfatizamos a importância de uma definição clara e precisa da *população de pesquisa*. No entanto, ao estudar *Amostragem*, o maior interesse está voltado para o problema de estimar ou inferir certas quantidades ou parâmetros de diversas características (variáveis) numéricas que podem ser medidas ou observadas para cada unidade da população. No caso de características ou variáveis categóricas, podem ser criadas variáveis numéricas indicadoras das categorias de resposta, tomando valor igual a *um* se a unidade é classificada na categoria em questão e valor igual a *zero* caso contrário. Desta forma, toda a teoria de amostragem se resume à estimação de parâmetros ou quantidades descritivas de variáveis numéricas que poderiam, em tese, ser medidas para todas as unidades da população de pesquisa.

De fato, cada característica numérica ou variável de interesse dá origem a um *vetor populacional*, que é o conjunto de valores da variável correspondentes às unidades da população. Por exemplo, se y é a variável de pesquisa (de interesse) e y_i é o valor dessa variável y para a unidade i, então $Y_U = \{y_1,\ y_2,\ ...,\ y_i,\ ...,\ y_N\}$ é o *vetor populacional* gerado pela variável y.

Pelo exposto, fica claro que a observação de várias variáveis sobre uma mesma população gera diversos vetores populacionais, cada um correspondendo a uma das variáveis observadas.

Em muitos casos, o interesse em estudar determinada população resume-se à necessidade de conhecer os valores de alguns *parâmetros* de uma ou mais variáveis que podem ser medidas ou observadas para unidades daquela população. Esses *parâmetros-alvo* (ou de interesse) podem ser quaisquer funções dos valores dos vetores populacionais. Os casos mais comuns ocorrem quando há interesse em estimar *totais*, *médias*, *proporções*, *razões*, *quantis* ou mesmo *variâncias*, *covariâncias* e *correlações*, sendo menos frequentes os casos de interesse por outros parâmetros.

Os principais *parâmetros* de interesse podem ser representados por meio das seguintes funções dos valores de variáveis na população.

O *total populacional* da variável y, que é definido por:

$$Y = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i \tag{2.1}$$

A *média populacional* da variável y, que é definida por:

$$\overline{Y} = \frac{Y}{N} = \frac{1}{N} \sum_{i \in U} y_i \tag{2.2}$$

Uma *proporção populacional* p é simplesmente a média populacional de uma variável y do tipo indicadora, que toma apenas os valores um ou zero para cada unidade. O total de uma variável desse tipo representa a contagem de unidades na população possuidoras do atributo de interesse e a média é exatamente essa contagem dividida pelo tamanho da população.

Um caso especial de proporção populacional de interesse ocorre com a definição da *função de distribuição cumulativa empírica populacional - FDCEP*, dada por:

$$F_y(a) = \frac{1}{N} \sum_{i \in U} I(y_i \leq a) \tag{2.3}$$

onde $a \in \mathbb{R}$ e $I(y_i \leq a)$ representa a função indicadora do evento $y_i \leq a$. A função $F_y(a)$ retorna a proporção de unidades na população U que têm valores de y menores ou iguais a a.

O *quantil populacional* q da distribuição da variável y é definido como o menor valor de y tal que a *FDCEP* tem valor maior ou igual a q, isto é:

$$T_y(q) = argmin\{F_y(a) \geq q\} \tag{2.4}$$

onde a função $argmin$ retorna o menor valor do argumento a da função $F_y(a)$ tal que a condição $F_y(a) \geq q$ é satisfeita. Por exemplo, a *mediana populacional* da variável y corresponde ao quantil obtido quando $q = 0,5$, isto é, $\text{Mediana}_y = T_y(0,5)$.

A *variância populacional* da variável y é dada por:

$$S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \overline{Y})^2 = \frac{1}{N-1} \left[\sum_{i \in U} {y_i}^2 - N\overline{Y}^2 \right] \tag{2.5}$$

O *desvio padrão - DP populacional* da variável y é dado por:

$$DP_y = S_y = \sqrt{S_y^2} \tag{2.6}$$

O *coeficiente de variação - CV populacional* da variável y é dado por:

$$CV_y = \frac{DP_y}{\overline{Y}} = \frac{S_y}{\overline{Y}} \tag{2.7}$$

Seja z outra variável de pesquisa, tomando valores z_i, $i \in U$. Define-se então a *razão de totais* das variáveis y e z como:

$$R = \frac{\sum_{i \in U} y_i}{\sum_{i \in U} z_i} = Y/Z \tag{2.8}$$

Define-se também a *covariância populacional* e a *correlação populacional* das variáveis y e z como:

$$S_{yz} = \frac{1}{N-1}\sum_{i\in U}(y_i - \overline{Y})(z_i - \overline{Z}) = \frac{1}{N-1}\left[\sum_{i\in U} y_i z_i - N\,\overline{Y}\,\overline{Z}\right] \tag{2.9}$$

e

$$\rho_{yz} = \frac{S_{yz}}{S_y S_z} \tag{2.10}$$

Até agora foram apresentadas as definições de alguns *parâmetros da população* que se deseja conhecer. No entanto, para conhecer exatamente o valor de qualquer dos parâmetros definidos, seria necessário conhecer **todos** os valores da variável (ou variáveis) de pesquisa naquela população, o que só seria possível mediante a realização de um *Censo* no qual a(s) variável(is) de pesquisa fosse(m) medida(s) ou observada(s) para todas as unidades da população.

Por outro lado, pode ser que estimativas desses parâmetros, com margens de erro conhecidas e controladas, sirvam para os propósitos dos interessados. Nesse caso, uma *pesquisa por amostragem* poderia resolver o problema com vantagens em relação a um *Censo*. Entre as vantagens mais diretas de pesquisas por amostragem podem ser mencionados os menores custos de obtenção das informações de interesse, a maior rapidez para obtenção dos resultados e a redução da carga de coleta de informações sobre a população de pesquisa.

De agora em diante, consideramos o caso em que a situação enfrentada é tal que basta conhecer *estimativas* dos parâmetros de interesse, bem como indicações da margem de erro a que tais estimativas estão sujeitas. A seguir, tomando por base a ideia de obter *estimativas* dos parâmetros de interesse, são apresentados os principais conceitos relacionados com a *amostragem de populações finitas*.

2.2 Amostra

Uma *amostra* $s = \{i_1, i_2, \ldots, i_n\}$ é qualquer subconjunto não vazio de unidades selecionadas da população $U (s \subset U)$ para observação visando *estimar* os parâmetros de interesse. Uma amostra de tamanho n é uma amostra contendo n *unidades* selecionadas da população U, sendo $1 \leq n \leq N$.

A notação $i \in s$ designa que a unidade i foi incluída na amostra s. A notação $s \ni i$ indica que a amostra s contém a unidade i. Quando escrevemos $\sum_{i\in s}$ estamos somando em i sobre o conjunto de rótulos de unidades incluídas na amostra s. Quando escrevemos $\sum_{s\ni i}$ estamos somando em s sobre o conjunto de amostras possíveis que contêm a unidade populacional i.

No contexto deste livro, apresentamos somente a teoria e resultados aplicáveis a *amostras probabilísticas*, isto é, a amostras selecionadas com base em regras de aleatorização bem definidas e que satisfazem as condições 1 a 3 enunciadas na Seção 1.3 e descritas de maneira mais formal na próxima seção.

Os dados amostrais para a variável y são representados por $Y_s = \{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\}$.

2.3 Amostragem probabilística

A *amostragem probabilística* é qualquer procedimento de amostragem que satisfaça todas as condições enumeradas a seguir:

1. O *espaço amostral* S, correspondente ao conjunto de todas as amostras s possíveis, é bem definido e poderia ser enumerado, ao menos teoricamente.
2. Uma probabilidade $p(s)$ conhecida (ou calculável) é associada a cada amostra $s \in S$, de tal modo que $\sum_{s\in S} p(s) = 1$. A função $p(s)$ é denominada *plano amostral*.
3. Uma única amostra s ($s \in S$) é selecionada para observação usando um mecanismo de aleatorização (sorteio) tal que a amostra s é escolhida com probabilidade igual a $p(s)$.
4. Cada unidade $i \in U$ tem uma probabilidade positiva de ser selecionada para a amostra, isto é: $\pi_i = P(i \in s) = \sum_{s\ni i} p(s) > 0, \forall\, i \in U$. A probabilidade π_i é denominada *probabilidade de inclusão* (de primeira ordem) da unidade i.

5. As probabilidades de inclusão das unidades selecionadas para a amostra e outros aspectos da estrutura do plano amostral são levados em conta ao fazer inferência sobre os parâmetros populacionais.

2.4 Estatísticas, estimadores e estimativas

Uma *estatística* é uma função real dos valores observados numa amostra da população, isto é, é qualquer função real $f(y_{i_1}, y_{i_2}, \ldots, y_{i_n})$. Considere os dois exemplos a seguir:

o *total amostral* ou *soma amostral* da variável y:

$$t(s) = t = \sum_{i \in s} y_i \tag{2.11}$$

a *média amostral* da variável y:

$$\overline{y} = \frac{t(s)}{n} = \frac{1}{n} \sum_{i \in s} y_i \tag{2.12}$$

Um *estimador* $\widehat{\theta}(s)$ é uma estatística usada para estimar um certo parâmetro θ de interesse. Antes de observarmos a amostra s, um estimador é uma variável aleatória cuja distribuição temos interesse de conhecer, pois dela dependem propriedades importantes do estimador. Por simplicidade, daqui para a frente vamos usar a notação $\widehat{\theta}$ para designar estimadores, sem explicitar sua dependência da determinação da amostra s, sempre que isso for possível.

Após a determinação da amostra s e a coleta dos dados das unidades nessa amostra, o valor calculado (observado) do estimador é chamado de *estimativa* do parâmetro.

Uma questão central da teoria da amostragem é como escolher bons estimadores para os parâmetros de interesse. Intuitivamente, bons estimadores seriam estatísticas cujos valores fiquem próximos do valor do parâmetro que buscam estimar. Para ajudar com essa questão, é essencial dispor de critérios para a escolha de estimadores. Mas antes disso, é útil definir algumas propriedades de estimadores que serão consideradas na elaboração de critérios de decisão ou escolha que vamos propor usar.

O *valor esperado* de um estimador $\widehat{\theta}$ é denotado por $E_p(\widehat{\theta})$. A notação $E_p(\bullet)$ designa o valor esperado da quantidade sob a distribuição de probabilidades induzida pelo plano amostral, isto é:

$$E_p(\widehat{\theta}) = \sum_{s \in S} \widehat{\theta}(s) p(s) \tag{2.13}$$

O *vício* (ou *viés* ou *tendência*) do estimador $\widehat{\theta}$ é definido como:

$$B_p(\widehat{\theta}) = E_p(\widehat{\theta}) - \theta \tag{2.14}$$

Algumas vezes é de interesse expressar o vício em termos relativos e se utiliza então o *vício relativo* do estimador, definido como:

$$RB_p(\widehat{\theta}) = \frac{B_p(\widehat{\theta})}{\theta} \tag{2.15}$$

Vício é uma característica indesejada num estimador, pois significa que a distribuição do estimador $\widehat{\theta}$ não é centrada no alvo de inferência θ. Diz-se que o estimador $\widehat{\theta}$ é *não viciado* (ou *não enviesado* ou *não tendencioso*) para o parâmetro θ quando seu valor esperado é igual ao parâmetro θ, isto é, quando $E_p(\widehat{\theta}) = \theta$, ou alternativamente, quando $B_p(\widehat{\theta}) = RB_p(\widehat{\theta}) = 0$.

Nosso primeiro critério para apoiar a escolha de estimadores sugere então que tratemos de usar *estimadores* sem vício, ou *não viciados*, ou ao menos *aproximadamente não viciados*. Quando isto for possível, teremos estimadores cuja distribuição será centrada no alvo desejado da inferência.

A *variância* do estimador $\widehat{\theta}$ é definida como:

$$V_p(\widehat{\theta}) = \sum_{s \in S} [\widehat{\theta}(s) - E_p(\widehat{\theta})]^2 p(s) \tag{2.16}$$

Quando um estimador é *não viciado*, sua *variância* mede a dispersão da distribuição do estimador em torno do alvo de inferência θ. Duas medidas alternativas dessa dispersão que dependem da variância são o *desvio padrão - DP* do estimador (também designado *erro padrão*), dado por:

$$DP_p(\widehat{\theta}) = [V_p(\widehat{\theta})]^{1/2} \tag{2.17}$$

e o *coeficiente de variação - CV* do estimador, dado por:

$$CV_p(\widehat{\theta}) = \frac{DP_p(\widehat{\theta})}{\theta} \tag{2.18}$$

O *desvio padrão* mede a dispersão da distribuição do estimador em unidade de medida igual à usada na mensuração da variável de interesse e o *CV* expressa essa medida em termos relativos, o que pode facilitar a interpretação e a comparação em cenários onde as unidades de medida de diferentes parâmetros podem ser distintas, mas exista interesse em comparar a dispersão de estimadores desses parâmetros.

Quando um estimador $\widehat{\theta}$ é *viciado*, uma medida mais adequada da dispersão da distribuição do estimador em torno do alvo de inferência θ é o *erro quadrático médio - EQM* dado por:

$

$$EQM_p(\widehat{\theta}) = \sum_{s \in S} [\widehat{\theta}(s) - \theta]^2 p(s) \tag{2.19}$$

Versões análogas do desvio padrão e do coeficiente de variação adequadas ao caso de estimadores viciados são o *erro médio - EM* e o *erro relativo médio - ERM* definidos, respectivamente, como:

$$EM_p(\widehat{\theta}) = [EQM_p(\widehat{\theta})]^{1/2} \tag{2.20}$$

e

$$ERM_p(\widehat{\theta}) = \frac{EM_p(\widehat{\theta})}{\theta} \tag{2.21}$$

Um mesmo parâmetro pode ter mais de um estimador não viciado disponível. Precisamos então de um segundo critério para ajudar na escolha de estimadores. Nosso segundo critério vai usar a *variância*, no caso de estimadores exatamente não viciados, ou o *EQM* nos outros casos. Como se quer ter estimadores com os menores erros de estimação, o segundo critério é o de escolher sempre os estimadores com o menor *EQM*, ou com a menor variância quando forem não viciados.

No contexto da Amostragem, diferente do contexto usual da Inferência Estatística, não se estabelece uma distribuição de probabilidade (ou modelo) para os valores da variável y na amostra (ou na população). Além disso, em geral os parâmetros que se deseja estimar não são responsáveis pela especificação de uma tal distribuição de probabilidades (ou modelo). Como já indicado, em geral os parâmetros de interesse são definidos como funções dos valores (considerados fixos, mas desconhecidos) da variável y na população.

Por esse motivo, na Amostragem de populações finitas, não há um procedimento geral para gerar estimadores que sejam ótimos nalgum sentido, como é o caso do *método da máxima verossimilhança* no contexto usual da Inferência Estatística. Os princípios usados em Amostragem para derivar estimadores dos parâmetros de interesse são baseados na simplicidade e no *método dos momentos*, como vamos ilustrar.

Suponha que o parâmetro-alvo é o *total populacional* Y. Nesse caso, o objetivo principal seria usar os *dados amostrais* $\{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\}$ para *estimar* $Y = \sum_{i \in U} y_i$.

Um segundo objetivo seria conseguir medir ou estimar também a *precisão* ou a *margem de erro* da estimativa produzida para Y.

Um *estimador linear* $\widehat{Y}_w$ do total populacional Y é uma combinação linear dos valores amostrais y_i com *pesos amostrais* w_i a serem definidos, isto é:

$$\widehat{Y}_w = \sum_{i \in s} w_i y_i \tag{2.22}$$

Podemos então usar os critérios sugeridos para escolha de estimadores para determinar os pesos w_i, como apresentado adiante.

Para ajudar a consolidar as ideias já apresentadas até aqui, fazemos agora uso de um exemplo muito simples, mas através do qual podemos ilustrar como operar com os conceitos e definições já introduzidos.

Exemplo 2.1. Considere os dados da Tabela 2.2 de uma população fictícia com $N = 4$ mulheres (unidades populacionais), de quem foi indagado o número de filhos tidos nascidos vivos (a variável y).

Valor da variável y por unidade da população de mulheres

Rótulo da unidade (i)	1	2	3	4	Total
Valor da variável (y_i)	0	0	2	1	3

Tabela 2.2: Valor da variável y por unidade da população de mulheres

Rótulo da unidade (i)	1	2	3	4	Total
Valor da variável (y_i)	0	0	2	1	3

Existem $\binom{4}{2} = 6$ amostras possíveis de duas unidades distintas dessa população, isto é, de tamanho $n = 2$. O conjunto de todas as amostras possíveis é dado por: $S = \{(1;2); (1;3); (1;4); (2;3); (2;4); (3;4)\}$.

A notação para representar o conjunto que forma cada amostra foi o ( ), para evitar usar { } dentro de { }. Cada elemento do conjunto S é, em si mesmo, um conjunto (neste caso, um par) de rótulos de unidades selecionadas para a amostra.

Considere agora um *plano amostral* p_1 em que uma qualquer das amostras possíveis é selecionada com igual probabilidade atribuída a todas as amostras possíveis. Considerando a condição 2, cada uma das seis amostras possíveis terá probabilidade igual a 1/6 de ser selecionada, isto é: $p_1(s) = 1/6, \ \forall\, s \in S$.

A Tabela 2.4 apresenta o conjunto de todas as amostras possíveis, os rótulos das unidades incluídas em cada amostra, os valores de y para as unidades incluídas na amostra, a soma amostral e as probabilidades de seleção de cada amostra. As colunas 1, 2 e 5 dessa tabela correspondem à apresentação detalhada do *plano amostral* p_1 tal como definido acima, agora representado na forma de uma tabela.

Informações de cada amostra possível sob plano amostral p_1

Amostra	Unidades na Amostra s	Valores na Amostra s	Soma Amostral (t)	Probabilidades $p_1(s)$
1	{1;2}	{0;0}	0	1/6
2	{1;3}	{0;2}	2	1/6
3	{1;4}	{0;1}	1	1/6
4	{2;3}	{0;2}	2	1/6
5	{2;4}	{0;1}	1	1/6
6	{3;4}	{2;1}	3	1/6
Total	—	—	—	1

A distribuição de probabilidades da estatística *Soma Amostral*, apresentada na Tabela 2.4, pode ser calculada a partir das informações na Tabela 2.4 e é dada por:

Probabilidade sob p_1 para cada valor de t

Valores possíveis de t	0	1	2	3
Com probabilidade $p_1(s)$	1/6	2/6	2/6	1/6

O valor esperado de t é:

$$E_{p_1}(t) = \sum_{s \in S} t(s)\, p_1(s) = 0 \times \frac{1}{6} + 1 \times \frac{2}{6} + 2 \times \frac{2}{6} + 3 \times \frac{1}{6} = \frac{9}{6} = 1,5$$

Porém o *total populacional* é $Y = \sum_{i \in U} y_i = 3$.

Como $1,5 = E_{p_1}(t) \neq Y = 3$, dizemos que t seria um *estimador viciado* de Y sob o plano amostral p_1 adotado.

Como poderíamos "corrigir" o estimador de modo que ficasse *não viciado* para o total populacional? Uma ideia simples vem da constatação de que $Y/E_{p_1}(t) = 3/1,5 = 2$. Logo, multiplicando por 2 o valor da soma amostral t resultaria num estimador cujo valor esperado deve ser igual a Y.

Considere então o novo estimador do total populacional dado por: $\widehat{Y} = 2 \times t$. Tal estimador pode ser escrito na forma linear como: $\widehat{Y} = 2 \times t = \sum_{i \in s} 2 \times y_i = \widehat{Y}_w$.

A Tabela 2.4 apresenta os valores possíveis e a distribuição de probabilidades do novo estimador $\widehat{Y}_w = 2 \times t$.

Probabilidade sob p_1 para cada valor de $2t$

Valores possíveis de $2t$	0	2	4	6
Com probabilidade $p_1(s)$	1/6	2/6	2/6	1/6

Verifica-se então que o valor esperado de $\widehat{Y}_w = 2 \times t$ é:

$$E_{p_1}(\widehat{Y}_w) = \sum_{s \in S} \widehat{Y}_w(s)\, p_1(s) = 0 \times \frac{1}{6} + 2 \times \frac{2}{6} + 4 \times \frac{2}{6} + 6 \times \frac{1}{6} = \frac{18}{6} = 3$$

Como agora $E_{p_1}(\widehat{Y}_w) = 3 = Y$, dizemos que $\widehat{Y}_w = 2 \times t$ é um *estimador não viciado* de Y sob o plano amostral p_1 considerado.

O método pelo qual deduzimos os pesos para usar com o estimador $\widehat{Y}_w$ não é viável na prática, pois foi preciso conhecer todos os valores da variável de pesquisa para obter o valor esperado do estimador inicialmente considerado (soma amostral) e só então calcular pesos que levariam à obtenção do estimador ponderado não viciado. Essa limitação é resolvida na próxima seção, onde apresentamos um método geral para obter pesos amostrais que levam sempre à estimação não viciada do total populacional.

2.5 A distribuição de aleatorização

A função $p(s)$ definida no conjunto S de todas as amostras possíveis é uma distribuição de probabilidades. Induzida por esta distribuição, é possível obter a distribuição de probabilidades de qualquer estatística (ou estimador) que seria calculada a partir dos dados coletados na amostra selecionada s. A distribuição de probabilidades assim obtida é chamada de *distribuição de aleatorização* da estatística ou estimador. Este foi o conceito ilustrado quando obtivemos a distribuição de probabilidades da estatística *soma amostral* no Exemplo 2.7.

Na *amostragem probabilística*, inferências são feitas considerando a *distribuição de aleatorização*. Tais inferências consideram como única fonte de variação ou incerteza a possível repetição hipotética do processo de amostragem utilizando o *plano amostral* $p(s)$, que resultaria em diferentes amostras $s_1, s_2, ... \in S$.

A distribuição de $\widehat{Y}_w = 2 \times t = \sum_{i \in s} 2 \times y_i$ determinada por $p(s)$ é também chamada de *distribuição amostral* do estimador. Vamos estudar suas propriedades para avaliar se $\widehat{Y}_w$ é um bom estimador para o total populacional Y.

2.6 Estimadores não viciados para o total populacional

No Exemplo 2.7 mostrou-se como se pode obter a *distribuição amostral* de um estimador (ou de uma estatística qualquer) a partir da distribuição de probabilidades induzida pelo plano amostral $p(s)$. Isto foi muito fácil de fazer porque contamos com duas condições favoráveis, que não se repetem na prática:

1) Os tamanhos da população N e da amostra n eram muito pequenos (4 e 2, respectivamente).

2) Consideramos conhecidos os valores da variável y para todas as unidades da população U.

Na grande maioria das situações de interesse prático no campo das pesquisas por amostragem, os tamanhos de população e amostra serão muito maiores. Também não serão conhecidos os valores da variável de interesse para unidades que não sejam selecionadas para a amostra que vai ser pesquisada. Num cenário com estas características, trabalhar com a distribuição $p(s)$ para daí tentar derivar distribuições amostrais de estimadores é complicado.

O primeiro problema é que o número total de amostras possíveis cresce muito rapidamente com N e com n. Por exemplo, o número de amostras sem reposição de tamanho n de uma população com N unidades é $\binom{N}{n}$. A Tabela 2.6 mostra como cresce o número de amostras no conjunto S para valores selecionados de N e n. Note como o tamanho desse conjunto é gigantesco mesmo com tamanhos de população e amostra bem modestos (1.000 e 20), por exemplo.

Tamanhos do espaço amostral S para valores selecionados de N e n

N	n	$\binom{N}{n}$
4	2	6
10	4	210
100	10	17.310.309.456.440
1.000	20	3,39482811302458e+41
1.0000	100	6,52084692454763e+241

Uma saída é usar propriedades simplificadoras da distribuição induzida pelo plano amostral. Tratamos disso na próxima seção, mas antes disso, vamos usar uma propriedade importante que pode ser deduzida a partir da distribuição de aleatorização.

A *probabilidade de inclusão* da unidade i na amostra é dada por:

$$P(i \in s) = \pi_i = \sum_{s \ni i} p(s) \tag{2.23}$$

Se tomarmos o *inverso da probabilidade de inclusão* $1/\pi_i$ como peso (w_i) de uma unidade amostrada, é fácil verificar que o estimador dado por $\widehat{Y}_w$ é *não viciado* para o total populacional Y:

$$\widehat{Y}_w = \sum_{i \in s} w_i y_i = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i \in s} {\pi_i}^{-1} y_i \tag{2.24}$$

Essa é propriedade importante e é demonstrada de maneira formal na próxima seção. Mas antes disso, vamos verificar sua aplicação com os dados do Exemplo 2.7. Continuando a discussão desse exemplo com a população de $N = 4$ mulheres de quem foi indagado o número de filhos tidos nascidos vivos (y), tem-se, na Tabela 2.6, o valor da variável y e a probabilidade de inclusão π_i de cada unidade da população de mulheres.

Valor da variável y e probabilidade de inclusão para cada unidade da população

Rótulo da unidade (i)	1	2	3	4	Total
Valor y_i	0	0	2	1	3
Probabilidade de inclusão π_i	3/6=1/2	3/6=1/2	3/6=1/2	3/6=1/2	-

Usando a propriedade recém apresentada, os pesos amostrais no Exemplo 2.7 seriam dados por $w_i = 1/\pi_i = \frac{1}{1/2} = 2$ para qualquer uma das unidades da população que fossem selecionadas para uma das amostras de tamanho $n = 2$.

O estimador ponderado do total nesse caso seria dado por:

$$\widehat{Y}_w = \sum_{i \in s} w_i y_i = \sum_{i \in s} {\pi_i}^{-1} y_i = \sum_{i \in s} 2y_i = 2t$$

e já se mostrou que este estimador é não viciado para Y.

Exemplo 2.2. Considere a mesma população fictícia do exemplo anterior. Considere agora o plano amostral p_2, que retira amostras de tamanho $n = 2$ dessa população com as probabilidades indicadas na Tabela 2.6.

Informações de cada amostra possível sob plano amostral p_2

Amostra	Unidades na Amostra s	Valores na Amostra s	Soma Amostral (t)	Probabilidades $p_2(s)$
1	{1;2}	{0;0}	0	0,00
2	{1;3}	{0;2}	2	0,20
3	{1;4}	{0;1}	1	0,15
4	{2;3}	{0;2}	2	0,20
5	{2;4}	{0;1}	1	0,15
6	{3;4}	{2;1}	3	0,30
Total	—	—	—	1,00

Vamos agora usar as informações acima para:

1. Verificar que a estatística soma amostral (t) é viciada para estimar o total populacional Y.
2. Obter / definir um estimador não viciado para o total populacional Y.

A distribuição da soma amostral t sob o plano p_2 é apresentada na Tabela 2.6.

Probabilidade sob p_2 para cada valor de t

Valores possíveis de t	0	1	2	3
Com probabilidade $p_2(s)$	0,0	0,3	0,4	0,3

O valor esperado de t sob o plano amostral p_2 é:

$$E_{p_2}(t) = \sum_{s \in S} t(s)\, p_2(s) = 0 \times 0,0 + 1 \times 0,3 + 2 \times 0,4 + 3 \times 0,3 = 2 < 3 = Y$$

Para obter um estimador não viciado, devemos calcular pesos adequados para as unidades amostrais. Estes requerem calcular as probabilidades de inclusão na amostra. A seguir, são apresentadas, na Tabela 2.6, as probabilidades de inclusão de cada uma das unidades da população e também os pesos amostrais correspondentes sob o plano p_2.

Probabilidade de inclusão e peso amostral de cada unidade sob o plano amostral p_2

Rótulo da unidade (i)	1	2	3	4
Probabilidade de inclusão (π_i)	7/20=0,35	7/20=0,35	7/10=0,70	3/5=0,60
Peso (w_i)	20/7=2,857	20/7=2,857	10/7=1,429	5/3=1,667

Usando o estimador do total com os pesos adequados $\widehat{Y}_w$, obtêm-se os valores das estimativas para cada amostra possível na coluna cinco da Tabela 2.6.

Obtenção de estimativa sob plano amostral p_2 para cada amostra possível

Amostra	Valores na Amostra s	Total Amostral ponderado	Probabilidades $p_2(s)$	Total × probabilidade
1	{0;0}	0	0,00	0
2	{0;2}	2x(10/7)	0,20	4/7
3	{0;1}	1x(5/3)	0,15	1/4
4	{0;2}	2x(10/7)	0,20	4/7
5	{0;1}	1x(5/3)	0,15	1/4
6	{2;1}	2x(10/7)+1x(5/3)	0,30	6/7+1/2
Total	-	-	1,00	3

‘‘‘

Notas:

1. O estimador $\widehat{Y}_w$ obtido usando os pesos iguais a $1/\pi_i$ tem valor esperado (ver total da última coluna da Tabela 2.6) igual ao total populacional Y. Logo, o estimador assim obtido é *não viciado* também sob o plano amostral p_2.
2. O fato de que a amostra 1 (composta pelas unidades {1;2}) tem probabilidade nula de ser selecionada não viola os critérios definidos para que o plano amostral p_2 seja chamado de *amostragem probabilística*. É fácil verificar que todas as condições enumeradas para que uma amostra seja declarada probabilística são cumpridas para esse plano amostral. Em particular, verifica-se que todas as unidades populacionais têm probabilidades positivas de inclusão na amostra - ver a Tabela 2.6.
3. Temos agora duas opções de plano amostral para selecionar amostras (de tamanho $n = 2$) da população U, visando estimar o total populacional Y. Com ambos os planos amostrais estão disponíveis estimadores não viciados do total populacional. Coloca-se então a pergunta: qual dos dois planos é melhor?

Estratégia 1: Seleção equiprovável de amostras com estimador de total ponderado ($\widehat{Y}_w = 2t$), conforme especificado na Tabela 2.6.

Probabilidade de seleção sob $p_1(s)$ para cada valor do estimador ponderado

Valores possíveis de $\widehat{Y}_w = 2t$	0	2	4	6
Com probabilidade $p(s)$ sob $p_1(s)$	1/6	2/6	2/6	1/6

Estratégia 2: Seleção de amostras com probabilidades desiguais e estimador de total ponderado ($\widehat{Y}_w$), conforme especificado na Tabela 2.6.

Probabilidade de seleção sob $p_2(s)$ para cada valor do estimador ponderado

Valores possíveis de $\widehat{Y}_w$	5/3	20/7	20/7+5/3
Com probabilidade $p(s)$ sob $p_2(s)$	0,30	0,40	0,30

A melhor estratégia é escolhida medindo o *afastamento esperado* entre os valores possíveis do estimador e o valor do total populacional desconhecido (Y). Para isso, como em ambos os casos o estimador é não viciado, usamos a *variância do estimador*. A Tabela 2.6 indica como pode ser calculada a variância de cada um dos estimadores sob as duas opções de plano amostral (p_2 e p_1).

Obtenção da variância dos estimadores sob os planos amostrais p_2 e p_1

Amostra	Valores na Amostra s	Estimativa sob p_2	Probabilidade sob p_2	Estimativa sob p_1	Probabilidade sob p_1
1	{0;0}	0	0,00	0	1/6
2	{0;2}	2x(10/7)	0,20	4	1/6
3	{0;1}	1x(5/3)	0,15	2	1/6
4	{0;2}	2x(10/7)	0,20	4	1/6
5	{0;1}	1x(5/3)	0,15	2	1/6
6	{2;1}	2x(10/7)+ 1x(5/3)	0,30	6	1/6
Variância	-	1,24	-	3,67	-

O plano amostral p_2 fornece o *estimador não viciado com menor variância* em comparação com o plano p_1 e deve ser preferido, pois o tamanho das amostras (nossa medida de custo) é o mesmo.

Minimizar a variância é o critério de desempate para escolha entre *estratégias não viciadas de amostragem e estimação* de *igual custo total*. Este será então nosso segundo critério para escolha de estimadores.

2.7 Teoria básica

Nesta seção, seguimos de perto a notação e a forma de apresentar os resultados encontrada no excelente livro de Särndal et al. (1992). Outra referência importante é o livro de Fuller (2009).

Como já foi dito, trabalhar com a distribuição $p(s)$ é complicado. Isto ocorre porque o número total, $\binom{N}{n}$, de amostras possíveis no conjunto S cresce muito rapidamente com N e com n. A saída encontrada é trabalhar com distribuições das variáveis aleatórias indicadoras δ_1, δ_2,... ,δ_N definidas tal que:

$$\delta_i = I(i \in s) = \begin{cases} 1, & i \in s \\ 0, & i \notin s \end{cases} \quad \forall\, i \in U \tag{2.25}$$

A variável δ_i é indicadora do evento 'inclusão da unidade i na amostra s'.

(Continuação) Para $N = 4$ e $n = 2$, as seis amostras possíveis podem ser representadas pelas indicadoras conforme apresentado na Tabela 2.7.

Representação de cada amostra possível pelas variáveis indicadoras

Amostra	Unidades na Amostra s	δ_1	δ_2	δ_3	δ_4
1	{1;2}	1	1	0	0
2	{1;3}	1	0	1	0
3	{1;4}	1	0	0	1
4	{2;3}	0	1	1	0
5	{2;4}	0	1	0	1
6	{3;4}	0	0	1	1

Cada amostra fica univocamente determinada pelas variáveis indicadoras δ_1, δ_2,... ,δ_N correspondentes. As variáveis indicadoras dependem da amostra s, apesar de não termos indicado isto explicitamente em nossa notação.

As probabilidades de inclusão na amostra, denotadas π_i, podem ser vistas como:

$$\pi_i = P(i \in s) = \sum_{s \ni i} p(s) = P(\delta_i = 1) = E_p(\delta_i), \;\; \forall\, i \in U$$

As *probabilidades de inclusão* π_i são ditas de *primeira ordem*.

Precisamos também definir *probabilidades de inclusão de segunda ordem*, denotadas π_{ij}, dadas por:

$$\pi_{ij} = P[(i,j) \in s] = \sum_{s \ni (i,j)} p(s) = P(\delta_i \delta_j = 1) = E_p(\delta_i \delta_j), \quad \forall\, (i,j) \in U$$

Note que quando $i = j$, $\pi_{ij} = \pi_{ii} = \pi_i, \ \forall\, i \in U$.

Além da propriedade de valor esperado das variáveis aleatórias indicadoras δ_i, pode-se também deduzir que:

$$V_p(\delta_i) = \pi_i(1 - \pi_i)$$

$$COV_p(\delta_i, \delta_j) = \pi_{ij} - \pi_i \pi_j$$

Um método geral de prova em amostragem se baseia num uso inteligente das variáveis indicadoras δ_1, δ_2,... ,δ_N. Uma propriedade importante dessas variáveis indicadoras é que:

$$\sum_{i \in s} \delta_i = \sum_{i \in U} \delta_i$$

Segue também que:

$$\sum_{i \in s} y_i = \sum_{i \in s} \delta_i y_i = \sum_{i \in U} \delta_i y_i$$

Note que o truque é converter a soma amostral que, antes de selecionada a amostra, tem parcelas aleatórias, em uma soma na população, onde as parcelas são conhecidas mas dependem das variáveis aleatórias indicadoras δ_i.

2.7.1 Estimador linear de total

Considere que o total populacional $Y = \sum_{i \in U} y_i$ é o parâmetro alvo. Um *estimador linear* de Y é sempre da forma:

$$\widehat{Y}_w = \sum_{i \in s} w_i y_i = \sum_{i \in U} \delta_i w_i y_i \tag{2.26}$$

onde w_i é o *peso amostral* da unidade i.

Para que o estimador linear $\widehat{Y}_w$ de Y seja **sempre** não viciado, é preciso que:

$$E_p(\widehat{Y}_w) = Y \Leftrightarrow \sum_{i \in U} E_p(\delta_i) w_i y_i = \sum_{i \in U} y_i \Leftrightarrow \sum_{i \in U} \pi_i w_i y_i = \sum_{i \in U} y_i$$

Esta relação só será válida para quaisquer valores populacionais y_i da variável de pesquisa caso $\pi_i \times w_i = 1, \ \forall\, i \in U$.

Portanto a condição para que o estimador linear do total $\widehat{Y}_w = \sum_{i \in s} w_i y_i$ seja **sempre** não viciado é que os pesos amostrais das unidades selecionadas sejam iguais ao inverso das respectivas probabilidades de inclusão de primeira ordem, isto é: $w_i = {\pi_i}^{-1} = 1/\pi_i = d_i, \ \forall\, i \in U$.

Os pesos amostrais d_i são chamados de *pesos básicos* do plano amostral. Doravante usamos a notação d_i para os pesos básicos, pois mais adiante vamos introduzir outros pesos amostrais. A notação w_i é reservada para designar pesos genéricos que podem ser aplicados para a obtenção de estimadores (viciados ou não).

Com os pesos básicos d_i, o estimador não viciado de total fica dado pelo conhecido *estimador de Horvitz-Thompson* ou *estimador HT*:

$$\widehat{Y}_{HT} = \sum_{i \in s} d_i y_i = \sum_{i \in s} {\pi_i}^{-1} y_i = \sum_{i \in s} y_i / \pi_i \tag{2.27}$$

Este estimador foi proposto por Horvitz e Thompson (1952) e está definido para qualquer variável de pesquisa y e para qualquer *plano amostral probabilístico*, isto é, plano em que $\pi_i > 0\,, \forall\, i \in U$. É para permitir desfrutar

dessa vantagem de sempre dispor de ao menos um estimador não viciado para totais que esta é uma das condições necessárias para a *amostragem probabilística* de populações finitas. Note também que o estimador faz uso das probabilidades de inclusão implicadas pelo plano amostral $p(s)$ adotado, mas depende deste apenas através das probabilidades de inclusão de primeira ordem das unidades selecionadas para a amostra, uma condição geralmente simples de satisfazer na prática da pesquisa.

2.7.2 Propriedades do estimador de Horvitz-Thompson

O *estimador de Horvitz-Thompson* é *não viciado* para estimar o total, ou seja, $E_p(\widehat{Y}_{HT}) = Y$.

Prova:

$$E_p(\widehat{Y}_{HT}) = E_p\left[\sum_{i\in U}\delta_i y_i/\pi_i\right] = \sum_{i\in U}[E_p(\delta_i)y_i/\pi_i] = \sum_{i\in U} y_i = Y$$

Esta propriedade vale para qualquer população, variável de interesse y e plano amostral, desde que $\pi_i > 0\,,\ \forall\, i \in U$.

A variância do estimador Horvitz-Thompson para o total é dada por:

$$\begin{aligned} V_{HT}(\widehat{Y}_{HT}) &= \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) y_i y_j \\ &= \sum_{i\in U}\sum_{j\in U}\left(\frac{d_i d_j}{d_{ij}} - 1\right) y_i y_j \end{aligned}$$

onde $d_{ij} = 1/\pi_{ij}$.

Esta é a chamada forma de Horvitz-Thompson da variância. Existe uma outra forma para esta variância, que vamos conhecer mais adiante.

Prova:

$$\begin{aligned} V_{HT}(\widehat{Y}_{HT}) &= V_p\left(\sum_{i\in U}\delta_i\frac{1}{\pi_i}y_i\right) \\ &= \sum_{i\in U}\sum_{j\in U} COV_p(\delta_i,\delta_j)\left(\frac{y_i}{\pi_i}\right)\left(\frac{y_j}{\pi_j}\right) \\ &= \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right) \\ &= \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) y_i y_j \\ &= \sum_{i\in U}\sum_{j\in U}\left(\frac{d_i d_j}{d_{ij}} - 1\right) y_i y_j \end{aligned}$$

Um estimador não viciado da variância do estimador HT do total é dado por:

$$\begin{aligned} \widehat{V}_{HT}(\widehat{Y}_{HT}) &= \sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right) \\ &= \sum_{i\in s}\sum_{j\in s}(d_i d_j - d_{ij})\, y_i y_j \end{aligned} \tag{2.28}$$

Este estimador da variância foi obtido usando o princípio dos estimadores tipo Horvitz-Thompson do total, mas agora, como se tratava de estimar uma soma dupla na população, os pesos das parcelas nessa soma dependem das probabilidades de inclusão de segunda ordem, isto é, das probabilidades de inclusão dos pares de unidades.

Para que este estimador seja viável, o plano amostral empregado tem que satisfazer a condição adicional de que as probabilidades de inclusão π_{ij} sejam estritamente positivas, $\forall\, i \neq j \in U$.

Para planos amostrais de tamanho prefixado, uma forma alternativa para a variância do estimador HT do total populacional, equivalente a apresentada anteriormente, é dada pela expressão de Sen-Yates-Grundy a seguir - ver Yates e Grundy (1953) e Sen (1953).

$$\begin{aligned} V_{SYG}(\widehat{Y}_{HT}) &= \sum_{i\in U}\sum_{j>i}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \\ &= \sum_{i\in U}\sum_{j>i}(1/d_id_j - 1/d_{ij})\,(d_iy_i - d_jy_j)^2 \end{aligned} \tag{2.29}$$

Note a troca do sinal da diferença de probabilidades de inclusão em relação à expressão anterior.

Uma análise dessa expressão de variância nos dá uma indicação de quando pode ser vantajoso empregar probabilidades de inclusão distintas. A variância do estimador de total seria nula caso $\frac{y_i}{\pi_i} = \frac{y_j}{\pi_j}$, $\forall\, i \neq j \in U$. Isto só seria possível quando $\pi_i \propto y_i\,, \forall\, i \in U$, isto é, quando as probabilidades de inclusão fossem exatamente proporcionais aos valores da variável de interesse. Na prática, é impossível aplicar essa ideia já que os valores da variável de interesse são desconhecidos antes da seleção da amostra.

Entretanto, vemos no Capítulo **??** que esta ideia pode ser usada de forma aproximada fazendo as probabilidades de inclusão proporcionais a uma medida de tamanho cujos valores estejam disponíveis para todas as unidades da população U. Sempre que a medida de tamanho for positivamente correlacionada com a(s) variável(is) de interesse y, vemos que é possível tirar proveito da informação de tamanho para aplicar métodos de amostragem que levam a estimadores mais eficientes do total que no caso de planos amostrais com equiprobabilidade para amostras de tamanhos iguais.

Um estimador alternativo da variância do estimador HT do total, pode ser escrito como:

$$\begin{aligned} \widehat{V}_{SYG}(\widehat{Y}_{HT}) &= \sum_{i\in s}\sum_{j>i}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \\ &= \sum_{i\in s}\sum_{j>i}(d_{ij}/d_id_j - 1)\,(d_iy_i - d_jy_j)^2 \end{aligned} \tag{2.30}$$

O estimador $\widehat{V}_{SYG}(\widehat{Y}_{HT})$ foi motivado a partir da forma de Sen-Yates-Grundy para a variância do estimador HT do total. Tal estimador não coincide com o estimador de variância derivado a partir da expressão de Horvitz-Thompson apresentada anteriormente.

Seguem alguns comentários sobre estimação de totais e respectivas variâncias em *amostragem probabilística*:

1. Com amostras probabilísticas, é sempre possível estimar sem vício um total populacional usando uma soma amostral π-ponderada, isto é, o estimador HT do total.

2. Há expressões de variância disponíveis para permitir avaliar a qualidade do estimador de total sob distintas situações (população, variável) para qualquer plano amostral probabilístico.

3. A estimação de muitos outros parâmetros populacionais (tais como médias, proporções e razões) usa em grande medida os resultados aqui apresentados para a estimação de totais. Isso fica mais claro nos capítulos seguintes.

4. Pode-se derivar estimadores não viciados do total populacional e da variância do estimador HT de total para distintos planos amostrais como casos especiais da teoria geral aqui apresentada. Isso é conveniente, em particular, para a estimação de variâncias, cujas expressões gerais dependem de somas duplas que podem se tornar inconvenientes de calcular quando os tamanhos de amostra são grandes. As expressões apresentadas para cada um dos planos amostrais específicos são úteis porque permitem simplificar os cálculos da estimação de variâncias.

2.7.3 Estimação da média populacional

Quando o tamanho da população N é conhecido, o estimador "natural" da média populacional baseado no estimador HT do total é:

$$\overline{y}_{HT} = \widehat{Y}_{HT}/N = \frac{1}{N}\sum_{i\in s} d_i y_i = \sum_{i\in s} w_i^{HT} y_i \tag{2.31}$$

onde $w_i^{HT} = d_i/N$.

As expressões de variância e seu estimador não viciado seguem diretamente das anteriores mediante divisão por N^2, levando a:

$$\begin{aligned} V_{HT}\left(\overline{y}_{HT}\right) &= \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) y_i y_j \\ &= \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\left(\frac{d_i d_j}{d_{ij}} - 1\right) y_i y_j \end{aligned} \tag{2.32}$$

e

$$\widehat{V}_{HT}\left(\overline{y}_{HT}\right) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\left(d_i d_j - d_{ij}\right) y_i y_j \tag{2.33}$$

Expressões na forma Sen-Yates-Grundy podem ser obtidas de forma análoga.

Mesmo quando o tamanho N da população não for conhecido, ele pode ser estimado usando o estimador HT do total de uma variável de contagem tomando valor igual a 1 para todas as unidades da população, levando ao estimador:

$$\widehat{N}_{HT} = \sum_{i\in s} d_i$$

Usando esse estimador do tamanho da população no denominador, um estimador tipo razão para a média populacional é dado por:

$$\overline{y}^R = \widehat{Y}_{HT}/\widehat{N}_{HT} = \frac{\sum_{i\in s} d_i y_i}{\sum_{i\in s} d_i} = \sum_{i\in s} w_i^R y_i \tag{2.34}$$

onde $w_i^R = d_i/\sum_{j\in s} d_j$.

A variância desse estimador de média pode ser aproximada por:

$$V_{HT}(\overline{y}^R) \doteq \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i - \overline{Y}}{\pi_i}\right)\left(\frac{y_j - \overline{Y}}{\pi_j}\right) \tag{2.35}$$

Um estimador aproximadamente não viciado para essa variância é dado por:

$$\widehat{V}_{HT}(\overline{y}^R) = \frac{1}{\widehat{N}_{HT}^2}\sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i - \overline{y}^R}{\pi_i}\right)\left(\frac{y_j - \overline{y}^R}{\pi_j}\right) \tag{2.36}$$

Cabe registrar que para alguns planos amostrais, os dois estimadores são equivalentes, isto é, $\overline{y}^R = \overline{y}_{HT}$ porque $w_i^R = w_i^{HT}$. Porém, quando diferem, o *estimador de razão da média* é geralmente mais eficiente que o estimador HT. Uma outra propriedade atraente do estimador tipo razão da média é que ele é invariante sob transformações de locação, isto é, se tomarmos $z_i = y_i + A$, então $\overline{z}^R = \overline{y}^R + A$. Esta propriedade não se verifica para o estimador HT.

Em planos amostrais equiponderados, isto é, em que as probabilidades de inclusão π_i são todas iguais, os pesos w_i para estimação de médias ficam todos iguais a $1/n$ para ambos os estimadores (HT e de Razão). Esta é uma vantagem de planos deste tipo, pois a tarefa de estimação fica simplificada.

A Tabela 2.7.3 apresenta um resumo dos estimadores HT do total, média e respectivas variâncias.

Estimadores HT do total, média e respectivas variâncias

Estimador
$\widehat{Y}_{HT} = \sum_{i\in s} d_i y_i = \sum_{i\in s} y_i/\pi_i$
$\overline{y}_{HT} = \widehat{Y}_{HT}/N = \sum_{i\in s} d_i y_i/N = \sum_{i\in s} w_i^{HT} y_i$
$\overline{y}^R = \frac{\sum_{i\in s} d_i y_i}{\sum_{i\in s} d_i} = \sum_{i\in s} w_i^R y_i$
$\widehat{V}_{HT}(\widehat{Y}_{HT}) = \sum_{i\in s}\sum_{j\in s} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right)$
$\widehat{V}_{SYG}(\widehat{Y}_{HT}) = \sum_{i\in s}\sum_{j>i} \left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$
$\widehat{V}_{HT}(\overline{y}_{HT}) = \widehat{V}_{HT}(\widehat{Y}_{HT})/N^2$
$\widehat{V}_{HT}(\overline{y}^R) = \frac{1}{\widehat{N}_{HT}^2}\sum_{i\in s}\sum_{j\in s} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \left(\frac{y_i-\overline{y}^R}{\pi_i}\right)\left(\frac{y_j-\overline{y}^R}{\pi_j}\right)$

2.8 Exercícios

Considere a população com $N = 6$ domicílios listada com os respectivos valores de variáveis de interesse na Tabela 2.8.

Valores de variáveis de interesse para cada domicílio da população

Domicílio	Renda (R$)	Número de Moradores	Número de Trabalhadores
1	800	2	2
2	4.200	4	3
3	1.600	2	1
4	500	2	1
5	900	4	2
6	2.000	1	1
Total	10.000	15	10

Tratar das seguintes questões:

1. Para cada variável de interesse (Renda, Número de Moradores e Número de Trabalhadores), calcule os seguintes parâmetros populacionais: total, média e variância.

2. Liste o conjunto S de todas as amostras possíveis de tamanho 2 da população, considerando apenas amostras de unidades distintas.

3. Supondo que todas as amostras listadas no conjunto S são equiprováveis (Plano A), calcule:
   a. As probabilidades de inclusão das unidades.

   b. As probabilidades de inclusão dos pares de unidade.

   c. Os valores possíveis para o estimador Horvitz-Thompson do total populacional para a variável Renda.

   d. O valor esperado e a variância para o estimador Horvitz-Thompson do total populacional para a variável Renda.

4. Considere agora que o conjunto S é formado somente pelas amostras (1;2), (2;3), (2;4), (2;5) e (2;6), tendo cada uma delas probabilidade 1/5 de ser a amostra selecionada (Plano B). Repita os cálculos do item 3 para o novo plano amostral.

5. Faça gráficos dos valores possíveis do estimador de total sob os dois planos amostrais para comparar as respectivas distribuições.

6. Use os resultados obtidos em 3 e 4 para comparar os dois planos amostrais e indique qual deles seria preferível usar, caso fosse necessário amostrar duas unidades distintas da população ($n = 2$) para estimar o total da Renda. Justifique.

Referências

Backstrom, C. H. e Hursh-César, G. (1981). *Survey Research, second edition* (p. 436). John Wiley & Sons.

Biemer, P. e Lyberg, L. (2003). *Introduction to survey quality* (p. 402). John Wiley & Sons.

Chambers, R. L. e Clark, R. G. (2012). *An Introduction to Model-Based Survey Sampling with Applications* (p. 265). Oxford University Press.

Freitas, M. P. S. e Antonaci, G. (2014). *Sistema Integrado de Pesquisa Domiciliares: Amostra Mestra 2010 e Amostra da PNAD Contínua* (p. 43). Instituto Brasileiro de Geografia e Estatística - IBGE. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv86747.pdf (Acesso: set. 2020.)

Fuller, W. A. (2009). *Sampling Statistics* (p. 454). John Wiley & Sons.

Groves, R. M.; Fowler, F. J.; Couper, M. P.; Lepkowski, J. M.; Singer, E. e Tourangeau, R. (2009). *Survey Methodology, 2nd edition* (p. 461). John Wiley & Sons Inc.

GSBPM. (2013). *Generic Statistical Business Process Model: GSBPM: version 5.1* (p. 30). United Nations Economic Commission for Europe - UNECE. Disponível em: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0 (Acesso: set. 2020.)

Horvitz, D. G. e Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

IBGE. (2003). *Pesquisa da Economia Informal Urbana* (p. 158). Instituto Brasileiro de Geografia e Estatística - IBGE.

IBGE. (2004). *Pesquisa Nacional por Amostra de Domicílios 2004* (Vol. 25, p. 1–116). Instituto Brasileiro de Geografia e Estatística - IBGE.

IBGE. (2014). *Pesquisa Nacional por Amostra de Domicílios Contínua - Notas Metodológicas - volume 1* (p. 47). Instituto Brasileiro de Geografia e Estatística - IBGE. Disponível em: ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Notas_metodologicas/notas_metodologicas.pdf (Acesso: set. 2020.)

IBGE. (2016). *Metodologia do Censo Demográfico 2010, 2a edição* (p. 711). Instituto Brasileiro de Geografia e Estatística - IBGE. Disponível em: http://biblioteca.ibge.gov.br/visualizacao/livros/liv95987.pdf (Acesso: set. 2020.)

IBGE. (2019). *Quadro geográfico de referência para produção, análise e disseminação de estatísticas* (p. 178). Instituto Brasileiro de Geografia e Estatística - IBGE.

Little, R. J. A. e Rubin, D. B. (2002). *Statistical analysis with missing data, 2nd edition* (p. 278). John Wiley &

Sons.

Magalhães, M. N. (2006). *Probabilidade e Variáveis Aleatórias* (Segunda ed, p. 411). EDUSP.

Magalhães, M. N. e Lima, A. C. P. (2004). *Noções de Probabilidade e Estatística* (6a. edição, p. 392). EDUSP.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (p. 258). John Wiley & Sons.

Särndal, C. E.; Swensson, B. e Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data* (p. 430). Chapman & Hall / CRC.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, *5*, 119–127.

Statistics Canada. (2020). *Guide to the Labour Force Survey*. Disponível em: https://www150.statcan.gc.ca/n1/pub/71-543-g/71-543-g2020001-eng.htm (Acesso: set. 2020.)

Thompson, S. K. (2012). *Sampling, third edition* (p. 343). John Wiley & Sons.

Valliant, R.; Dorfman, A. H. e Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach* (p. 504). John Wiley & Sons.

Waal, T.; Pannekoek, J. e Scholtus, S. (2011). *Handbook of statistical data editing and imputation* (p. 439). John Wiley & Sons.

Wild, C. J. e Seber, G. A. F. (2004). *Encontros com o acaso: um primeiro curso de análise de dados e inferência* (p. 411). LTC - Livros Técnicos e Científicos Editora S.A.

Yates, F. e Grundy, P. M. (1953). Selection Without Replacement from Within Strata with Probability Proportional to Size. *Journal of the Royal Statistical Society. Series B (Methodological)*, *15*(2), 253–261.