Aula 05
Leitura Obrigatória

Política editorial do PROJETO FISHER

O PROJETO FISHER, uma inciativa da Associação Brasileira de Estatística, ABE, tem como finalidade publicar textos básicos de Estatística em língua portuguesa.

A concepção do projeto se fundamenta nas dificuldades encontradas por professores dos diversos programas de bacharelado em Estatística no Brasil em adotar textos para as disciplinas que ministram.

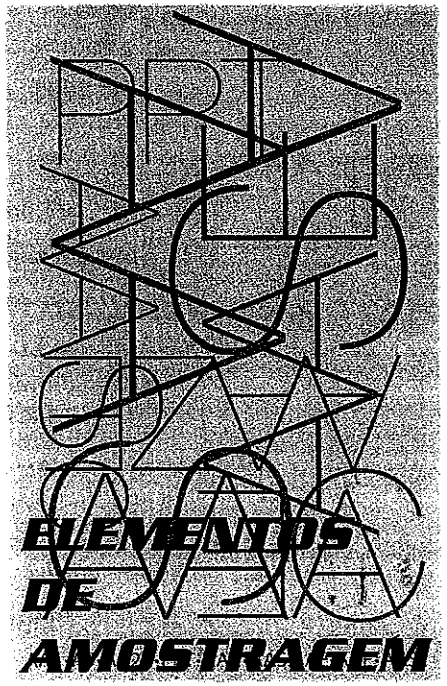
A inexistência de livros com as características mencionadas, aliada ao pequeno número de exemplares em outro idioma existente em nossas bibliotecas impedem a utilização de material bibliográfico de uma forma sistemática pelos alunos, gerando o hábito de acompanhamento das disciplinas exclusivamente de notas de aula.

Em particular, as áreas mais carentes são: Amostragem, Análise de Dados Categorizados, Análise Multivariada, Análise de Regressão, Análise de Sobrevivência, Controle de Qualidade, Estatística Bayesiana, Inferência Estatística, Planejamento de Experimentos etc.

Embora os textos que se pretende publicar possam servir para usuários da Estatística em geral, o foco deverá estar centrado nos alunos do bacharelado. Nesse contexto, os livros devem ser elaborados procurando manter um alto nível de motivação, clareza de exposição, utilização de exemplos preferencialmente originais e não devem prescindir do rigor formal. Além disso, devem conter um número suficiente de exercícios e referências bibliográficas e apresentar indicações sobre implementação computacional das técnicas abordadas.

A submissão de propostas para possível publicação deverá ser acompanhada de uma carta com informações sobre o objetivo de livro, conteúdo, comparação com outros textos, pré-requisitos necessários para sua leitura e disciplina onde o material foi testado.

ABE - Associação Brasileira de Estatística



Heleno Bolfarine
e
Wilton O. Bussab

Universidade de São Paulo
Instituto de Matemática e Estatística

ABE - PROJETO FISHER

EDITORA BLUCHER
www.blucher.com.br

tragem e dependendo do número de horas destinado ao curso, sugere-se usar como complemento o livro de Pessoa e Silva (1998).

Durante estes anos recebemos várias sugestões de professores, colegas e alunos as quais procuramos incorporar nesta versão, e afirmamos que erros persistentes são de nossa inteira responsabilidade. Sem o incentivo e cobrança das diretorias da ABE também não teríamos retomado a tarefa de transformar aquelas notas em livro. O bacharel em estatística Frederico Z. Poleto fez um excelente trabalho de revisão, digitação e sugestões para o aprimoramento do texto. Agradecemos a todos.

Capítulo 1

Noções básicas

A experiência com amostragem é fato corrente no cotidiano. Basta lembrar como um cozinheiro verifica o tempero de um prato que está preparando, como alguém testa a temperatura de um prato fumegante de sopa, ou ainda como um médico detecta as condições de um paciente através de exames de sangue. Poderiam ser listados outros exemplos que usam procedimentos amostrais mais complicados, mas todos com o mesmo objetivo: obter informações sobre o todo, baseando-se no resultado de uma amostra.

Porém, o uso inadequado de um procedimento amostral pode levar a um viés de interpretação do resultado. Por exemplo, não mexer bem a sopa antes de retirar uma colher para experimentar pode levar à subavaliação da temperatura do prato todo com conseqüências desagradáveis para o usuário.

Em estudos mais sofisticados, onde as informações são obtidas através de levantamentos amostrais, é comum o usuário ficar tão envolvido na apuração e interpretação dos dados que "esquece" de verificar possíveis viéses originários do protocolo de escolha da amostra.

O uso de amostras que produzam resultados confiáveis e livres de viéses é o desejo de todos. Entretanto, estes conceitos não são triviais e precisam ser estabelecidos para o uso científico dos processos amostrais. Desse modo, necessita-se de teoria que descreva as propriedades e impropriedades de alguns protocolos de obter amostras. Esse é o objetivo do livro: apresentar os princípios básicos de uma "Teoria de Amostragem". Cursos introdutórios de inferência estatística também ensinam a fornecer resultados para o todo, baseando-se em resultados da amostra, porém a ênfase é dada para populações infinitas, ou o que é muito mais comum, a amostra é retirada de uma distribuição de probabilidade. Não se discute muito como a amostra

é obtida, garante-se apenas que as observações foram obtidas independentemente, com igual probabilidade, e retiradas de uma mesma população teoricamente infinita. Aqui a população será finita, e possivelmente enumerável ou passível de descrição.

Neste capítulo, pretende-se dar uma visão geral das questões envolvidas em um plano amostral e que servirá para um primeiro contato com aspectos metodológicos emergentes de uma pesquisa de tal natureza.

1.1 Palavras-chave

Toda teoria, e amostragem não foge à regra, necessita de um conjunto de conceitos e termos técnicos sobre o qual ela se fundamenta. Estes conceitos irão aparecendo pelos diversos capítulos conforme se tornarem necessários. Porém, é conveniente para unificar a linguagem e tornar mais clara a explicação, definir alguns desses conceitos, mesmo que de forma abreviada. No Apêndice, A estão listadas e descritas algumas palavras-chave que atendem a esse objetivo. Recomendamos ao leitor consultá-lo sempre que tiver dúvidas em relação a algum dos conceitos mencionados.

1.2 Guia para um levantamento amostral

Ao optar por uma pesquisa quantitativa, levantamento ou experimentação, é necessário que o pesquisador planeje, execute, corrija e analise adequadamente o procedimento proposto e usado. Isto significa tomar uma série de medidas e cuidados antes da realização, durante a aplicação e depois da pesquisa efetuada. Sem esses passos, dificilmente pode-se garantir resultados convincentes e confiáveis. Um estatístico experiente desenvolve os seus próprios procedimentos, escritos ou não, para conduzir ou orientar uma pesquisa quantitativa, mas terá muita dificuldade em transmitir esses conhecimentos sem a prática e o convívio cotidiano com o aprendiz. Um dos métodos para transferir conhecimento e agilizar o treinamento nesta atividade é através da apresentação de uma lista de tópicos que devam ser abordados em uma pesquisa quantitativa, ou melhor, apresentando o chamado "checklist". Estas listas nunca são definitivas ou completas. Em primeiro lugar elas traduzem as idiossincrasias de seus formuladores e, em segundo, dificilmente conseguem prever todas as possíveis situações de um mundo tão rico e complexo como as pesquisas quantitativas. Portanto, devem ser usadas como um guia aproximado para planejamento e execução de um plano amostral.

Apresentamos no Apêndice B a nossa lista de pontos. Ela é resultante de nossas discussões, conhecimento, aprendizado, experiência e prática. Além de servir como referência, aproveitaremos a relação para abordar alguns tópicos que raramente aparecem em livros de técnicas de amostragem. Tais assuntos são fundamentais para aqueles que tenham que conduzir ou assessorar um levantamento amostral, e ousamos afirmar que, se estes procedimentos metodológicos não forem adequados, não existe técnica estatística, por melhor ou mais sofisticada que seja, que possa produzir resultados idôneos.

Embora exista alguma aparente ordem na seqüência das atividades, a prática nem sempre age deste modo. Salta-se de um ponto para outro de acordo com as necessidades, lembranças e informações que vão aparecendo. Entretanto, seguir os pontos mencionados terá a vantagem de uma apresentação aparentemente mais racional, servindo também como roteiro para apresentação do relatório.

As seções seguintes abordarão alguns dos itens mencionados, procurando explicar um pouco mais sobre o seu significado. Os assuntos não serão obrigatoriamente tratados nem na ordem nem no grupo onde apareceram mencionados. Os demais capítulos deste livro, relacionados com as técnicas de amostragem, abordam com maior profundidade os itens contidos no grupo intitulado **Planejamento e Seleção de Amostra**.

1.3 O que se pretende conhecer?

1.3.1 Qual a questão a ser respondida?

Usualmente, o objetivo geral de uma pesquisa é óbvio. Na maioria das vezes, pode ser resumido em uma pergunta. As dificuldades começam ao se procurar respostas a esta pergunta. Qual o potencial do mercado no município X para consumir um novo produto cultural? Deve-se investigar as pessoas mais ricas ou as de maior nível educacional? O conhecimento substantivo do assunto abordado ajuda muito a estabelecer os melhores caminhos em busca de uma resposta? Estudar levantamentos semelhantes realizados no passado, ou em outras regiões, é uma das melhores fontes para identificar e operacionalizar objetivos, bem como obter sugestões de como o problema pode ser resolvido. Pode-se aprender muito com erros cometidos por outros pesquisadores.

Portanto, uma das maiores dificuldades de qualquer pesquisa é a formulação correta dos seus objetivos gerais e operacionais. Exige muito conhecimento específico

da área de interesse, muito trabalho de pesquisa bibliográfica e grande habilidade criativa por parte dos pesquisadores envolvidos. Em pesquisas quantitativas, a situação agrava-se pela necessidade de transformar estes objetivos em questões operacionais quantificáveis. A literatura, e a experiência mais ainda, é rica em exemplos e situações onde a distância entre o objetivo genérico e a resposta quantitativa operacional é muito grande. Pense, por exemplo, na questão: renda é uma boa maneira de operacionalizar o conceito de classe social para uma família? Caso a resposta seja afirmativa, o que é melhor: renda familiar total ou renda familiar per capita?

Pode-se até postular que "um problema corretamente definido já está resolvido", pois em sua formulação vem embutida a solução.

Quase sempre um levantamento amostral tem múltiplos objetivos, mas para efeitos práticos é conveniente prender-se a um conjunto pequeno de questões-chave e que precisam ser respondidas. Isto facilitará o trabalho de planejamento. As demais questões farão parte de um conjunto de objetivos secundários, que poderão ou não ser adequadamente respondidos pela pesquisa. Deve-se evitar fortemente a tentação de acrescentar questões só para aproveitar o levantamento.

1.3.2 A operacionalização dos conceitos

Um dos maiores desafios das pesquisas quantitativas é a criação de bons indicadores (variáveis, escalas) que representem adequadamente os conceitos (constructos) de interesse. São exemplos de constructos: inteligência, nível sócio-econômico, desempenho escolar, potencial de mercado, ansiedade, satisfação, etc. Para inteligência é bem conhecido o quociente de inteligência (QI) como um indicador. O critério Brasil, antigo ABA/ABIPEME, aquele que combina grau educacional, condições da moradia e bens possuídos é muito usado para expressar o nível sócio-econômico. O Ministério da Educação aplica uma série de provas para avaliar desempenho escolar (SAEB, ENEM, Provão, Pisa, etc.). Já para o potencial de mercado, procura-se criar uma escala medindo as componentes do conceito operacional: "pessoas, com dinheiro e disponibilidade para gastar". Estas escalas, muitas vezes mal entendidas e erroneamente empregadas, são aceitas e largamente usadas por terem sido validadas, isto é, foram criadas, analisadas contextualmente, comparadas e verificada a pertinência entre os valores na escala e o significado dentro do conceito. Alguns indicadores são medidos por meio de uma única variável mensurável, outros, que é o mais comum, são combinações de resultados de várias perguntas quantificáveis. Boa parte dos conteúdos dos livros de metodologia de pesquisa dedica-se a prescrever métodos e processos para transformar conceitos teóricos em escalas confiáveis e validadas. Dentro da vasta literatura disponível, recomenda-se o livro de Pedhazur e Schmelkin (1991), pela sua abordagem mais quantitativa.

1.3.3 Variáveis e atributos

Associada a cada *unidade elementar* (UE - veja a definição na Seção 1.4.1) existirá uma ou mais características de interesse à pesquisa. São as chamadas variáveis ou atributos. Por exemplo, em um estudo onde a UE é a família pode-se estar interessado na renda familiar total, no número de membros, no sexo ou educação do chefe, etc. Já para a UE empresa, o interesse pode ser no faturamento total, lucratividade, ramo de atividade econômica, consumo de energia elétrica, etc.

O objetivo específico da pesquisa é que orienta a escolha e definição da UE e das variáveis a serem coletadas. Em pesquisa de Marketing, sobre o poder de compra, uma das variáveis mais usadas é a renda **familiar** total. Já para um estudo sobre política de emprego é mais indicado analisar a renda **individual** do trabalhador. Em algumas situações, a escolha da UE é muito mais complexa. Por exemplo, em um estudo sobre o comportamento de setores ligados à indústria de alimentação, como tratar o restaurante dentro de uma grande montadora de automóveis? Observe que dependendo da definição, o mesmo estabelecimento poderia ser tratado de modo diferente, caso a exploração fosse própria ou terceirizada.

1.3.4 Especificação dos parâmetros

Com os conceitos de interesse da pesquisa traduzidos em variáveis mensuráveis, necessita-se tornar bem claro quais as características populacionais (**parâmetros**) que deverão ser estimados pela amostra. A falta de uma inequívoca definição inicial tem sido fatal para muitas pesquisas.

Suponha-se que o objetivo de um levantamento seja medir o crescimento das vendas das empresas do setor de vestuário em um determinado ano. Isso pode ser medido, pelo menos, de duas maneiras: (i) como a média do crescimento de cada empresa (vendas deste ano/vendas do ano anterior, para cada empresa) ou, (ii) razão entre o total de vendas de todas as empresas neste ano dividido pelo total de vendas das empresas no ano passado. Estes resultados podem ser bem diferentes, principalmente se as grandes empresas tiverem comportamento distinto das pequenas. A escolha de um outro parâmetro é fundamental na orientação do desenho amostral.

Quando o levantamento exige, além de estimativas para a população toda, também para estratos e/ou subpopulações, deve-se redobrar o cuidado no planejamento para garantir estimadores adequados para o todo e as partes. É bom lembrar que podem ser usadas diferentes formas de parâmetros para variáveis em estratos distintos.

1.4 De quem se está falando

1.4.1 Unidade elementar, amostral e resposta

A **unidade elementar**, ou simplesmente elemento de uma população, é o objeto ou entidade portadora das informações que pretende-se coletar. Pode ser uma pessoa, família, domicílio, loja, empresa, estabelecimento, classe de alunos, escola, etc. É muito importante que a unidade elementar seja claramente definida, para que o processo de coleta e análise tenha sempre um significado preciso e uniforme. Por exemplo, o conceito de família parece ser "natural", mas, sem uma definição adequada pessoas distintas teriam dificuldade de dar uma mesma classificação para situações especiais. Veja um destes casos: suponha que em um domicílio vive um casal com filhos adultos, inclusive uma de suas filhas casada, com o genro e um neto. Deve-se considerar uma ou duas famílias? Suponha, agora, que a filha é divorciada, e claro, o genro não vive com eles: mudaria alguma coisa na sua definição? Nestas situações, em vez de tentar criar definições próprias, recomenda-se fortemente buscar estudos já realizados, onde esses problemas já foram estudados e as definições serão mais amplas e permitirão comparações entre diferentes pesquisas. Para o exemplo citado acima, sugere-se consultar os manuais de metodologia de pesquisa editados pelo IBGE.

Qualquer plano amostral fará recomendações para selecionar elementos da população por meio das **unidades amostrais.** Pode ser formado por uma única unidade elementar ou por várias. Uma pesquisa eleitoral usa eleitores como sendo a unidade elementar. Um levantamento pode escolher um ponto da cidade e entrevistar os cem primeiros eleitores que passam por lá. Usou-se a unidade elementar como unidade amostral. Um plano alternativo decidiu selecionar domicílios e entrevistar todos os eleitores residentes nos domicílios escolhidos. A unidade elementar continua sendo eleitor, mas agora a unidae amostral passou a ser domicílio, um conjunto de unidades elementares. Como será visto mais à frente, os planos amostrais em múltiplos estágios empregam diferentes unidades amostrais em um mesmo planejamento. Por exemplo, uma amostra de eleitores pode ser obtida selecionando primeiro algumas cidades, quateirões dentro das cidades, domicílios dentro dos quateirões e finalmente eleitores dentro dos domicílios.

Às vezes, é conveniente ressaltar quem é a unidade respondente ou a **unidade de resposta.** Um exemplo pode ajudar a entender o conceito. O censo demográfico tem uma primeira parte com questões simples sobre cada morador do domicílio, tais como sexo, idade, grau de instrução, etc. Um único morador pode responder por todos os outros; usualmente, elege-se o chefe, ou cônjuge, como unidade de resposta.

1.4.2 As diversas populações possíveis

Como já foi dito, o objetivo da amostragem é fazer afirmações sobre uma **população**, baseando-se no resultado (informação) de uma amostra. Assim, não se sabendo exatamente de onde foi retirada a amostra, não se sabe para quem pode-se estender as conclusões, ou seja, para que população pode ser feita a inferência.

Inicialmente convém lembrar que se entende por **população** a reunião de todas as unidades elementares definidas no item anterior.

Como no caso dos objetivos, começa-se falando de uma população genérica e freqüentemente óbvia. Por exemplo, na pesquisa de potencial de mercado mencionada acima, decide-se investigar a renda individual dos moradores do município. Portanto, a população é formada por todos os moradores do município. Será que os jovens irão consumir o produto? E os moradores da região rural? Assim, em uma segunda aproximação operacional, a população passa a ser os adultos (maiores de 18 anos), moradores da região urbana de X. Restam ainda outras dúvidas: como tratar os inativos e aqueles que não têm renda? Conforme a resposta, pode ser necessário redefinir a **população objetivo** (ou população-alvo).

A obtenção de uma amostra, qualquer que seja o plano amostral adotado, necessita de uma relação das unidades elementares. O ideal seria dispor de um rol seqüencial dessas unidades para que se pudesse fazer uma escolha conveniente das unidades que comporiam a amostra. Entretanto, raramente dispõe-se de tais listas. No exemplo acima, dever-se-ia dispor da relação dos moradores de X, o que parece ser bem pouco provável que exista. Felizmente, existem informações, mais ou menos atualizadas, que podem ser usadas como alternativas para (descrever) a relação das unidades. Podem ser mapas, várias listas que, reunidas, descrevem boa parte do universo, censos, etc. Essas fontes que descrevem o universo a ser investigado formam o chamado **sistema de referências.** As unidades que aparecem nessas

listas muitas vezes são chamadas de **unidades de listagem.**

Para o exemplo de potencial mencionado acima, pode-se usar como sistema de referência a relação dos Setores Censitários (SC) empregada pelo IBGE nos Censos Demográficos. O município é dividido em pequenas áreas que, reunidas, recobrem toda a área do município. Durante a realização do censo, cada SC é designado a um entrevistador que se encarrega de aplicar o questionário em todos os moradores de cada domicílio. Aos interessados, o IBGE fornece o mapa do SC, o número e tipo de domicílios existentes, o total de moradores e uma série de outras informações agregadas. Na região urbana, cada SC engloba cerca de 300 domicílios. Essas informações são atualizadas de 10 em 10 anos, e algumas vezes em prazos menores. Analisando-se a relação de SC do município X, observa-se que em alguns deles existem quartéis, internatos, alojamentos, etc., os chamados **domicílios coletivos.** Também constata-se que alguns SCs são formados especificamente por favelas e, neste momento, não interessaria ao levantamento. Decide-se, assim, não entrevistar os domicílios coletivos e nem as favelas. Informações recentes sobre o crescimento da cidade, desde a última atualização dos SCs, informa que a cidade já está invadindo SCs que são classificados como rurais, mas não se sabe quais. Assim, devido à particularidade do sistema de referência, a população que servirá de base para a escolha da amostra pode ser definida como: "todos os moradores adultos, com residência em domicílios particulares classificados no último censo como moradores de região urbana, excluindo moradores de favelas". Repare que a definição operacional baseada no sistema de referência não é obrigatoriamente a mesma que a população-alvo. Chamaremos esta de **população referenciada** ou população referida.

Selecionada a amostra, passa-se ao trabalho de campo, onde os dados serão coletados. Por diversas razões, não se conseguem informações sobre algumas unidades selecionadas, e em compensação aparecem dados para outras unidades que não estavam previstas inicialmente. Unidades inexistentes, recusas, domicílios vagos, ou fechados, impossibilidade de acessar a unidade (condomínios fechados) são alguns dos motivos para se perder unidades. Criação de novos conjuntos habitacionais, transformação de casas em cortiços, etc. podem ser motivos de aparecimento de unidades não selecionadas a priori. Em todo caso, tem-se uma amostra que foi retirada de uma população que não é exatamente a referida. Se a cidade tiver muitos condomínios fechados, aos quais não foi permitido o acesso, e sabendo-se que nestes locais moram pessoas de alta renda, a estimativa do potencial de mercado será subestimada. Assim, a inferência referir-se-á apenas a uma nova população: a **população amostrada.** Ela só pode ser descrita, após a realização do levantamento

de campo, e procura-se ressaltar quais as possíveis diferenças que ela possa ter com a população referida.

A Figura 1.1 procura ilustrar as relações existentes entre as diferentes populações. Como a amostra foi retirada da população amostrada, é apenas sobre ela que valem as inferências estatísticas. A análise qualitativa, e algumas vezes até a quantitativa, das características das unidades perdidas e das agregadas permite avaliar quais as conseqüências em estender estas conclusões para a população referida. O conhecimento substantivo do assunto de pesquisa e das características das unidades distintas nas duas populações permite ao pesquisador avaliar as conseqüências de usar as conclusões da população referida para a população-alvo.

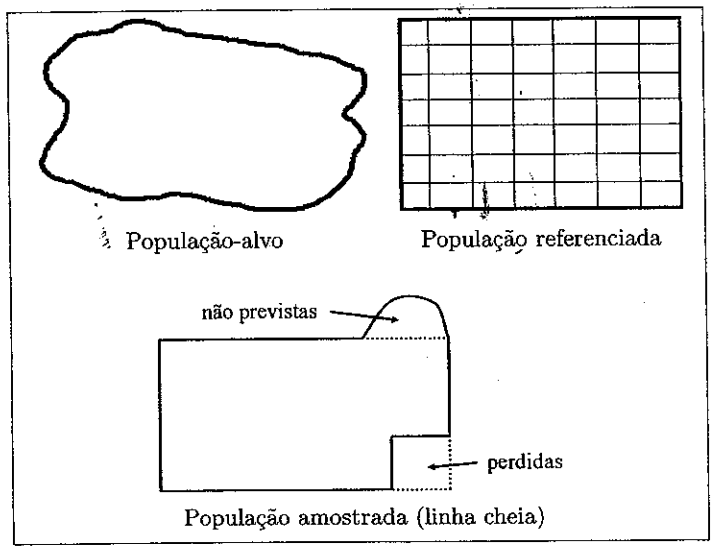


Figura 1.1: Comparações das populações-alvo, referenciada e amostrada

No exemplo em questão, estima-se estatisticamente qual o potencial relativo de pessoas na população amostrada. Para a população referida, pode-se apenas dizer que essa porcentagem deve ser maior que a da população amostrada. Não se saberia precisar o quanto, pois deixaram-se de lado informações desconhecidas sobre moradores mais ricos da cidade. Ao se eliminarem do sistema de referência as favelas e os domicílios coletivos, elimina-se também uma parte dos mais pobres. Se este contingente for maior que o dos moradores dos condomínios fechados, então o potencial relativo da população-alvo é menor do que o da população amostrada. Novamente, não se sabe precisar os valores do potencial sem outros estudos ou

informações.

Em sua opinião, e ainda usando o exemplo acima, de que modo a inclusão dos moradores rurais na população-alvo, afetaria o potencial de compra da cidade?

Caso a pesquisa deva produzir respostas para partes preestabelecidas da população, isto deve ser conhecido antes da definição do plano amostral. Suponha-se que no exemplo anterior pretendia-se conhecer o mercado potencial separado dos moradores das regiões sul e norte. Assim, antes de definir a amostra, devia-se separar o sistema de referências nos SCs do sul e do norte, ou seja, é como se estivesse trabalhando com duas populações. Cada uma dessas subpopulações é chamada de um **estrato**. Estratificação é uma das estratégias mais usadas em desenhos amostrais. É utilizada tanto para dar respostas a partes da população como para melhorar os processos de estimação. Será visto em outros capítulos como a estratificação é um recurso poderosíssimo dentro da Amostragem.

Existe uma forte tentação em usar a pesquisa amostral para conhecer detalhes de todas as partes da população, e para tanto, exagera-se no estabelecer o número de estratos. Esta opção freqüentemente implica em tamanhos de amostras economicamente inviáveis. Uma solução de compromisso é considerar os fatores básicos como estratos e os secundários como subclasses. Estas são partes da subpopulação que não entram no desenho amostral, mas são analisados a posteriori. Novamente, no exemplo em pauta, controla-se a amostra garantindo representantes do sul e do norte. Mas, pretende-se também conhecer o potencial segundo o sexo do respondente. Observe que, por não ter sido controlado o fator sexo, a amostra pode ter um número insignificante de representantes de uma das categorias de gênero, invalidando qualquer conclusão.

Solicita-se a atenção para a diferença entre estrato e subclasse. Ambos representam partes da população, porém o primeiro é contemplado no desenho amostral garantindo-se, a priori, estimativas confiáveis. Já na segunda, a qualidade das estimativas dependerá da presença ou não de unidades suficientes em cada subclasse. Maiores esclarecimentos sobre estas diferenças aparecerão nos capítulos técnicos.

Uma última palavra de advertência sobre os cuidados em definir as populações. Não se duvida em afirmar que o sucesso de um levantamento amostral baseia-se fortemente no conhecimento que se tem sobre a população. Deve-se gastar boa parte do tempo (mais de 50%) estudando e definindo a população. Dever-se-ia conhecer tanto sobre ela que talvez fosse até dispensável a realizacão da pesquisa.

1.5 Como obter os dados?

1.5.1 Tipos de investigação

Uma das etapas importantes de uma pesquisa quantitativa, e muitas vezes relegada a um segundo plano, é o levantamento dos dados da(s) característica(s) de interesse. Um exemplo bem conhecido de coleta de dados são os chamados censos populacionais, realizados no Brasil pelo IBGE, que procuram determinar o número de pessoas existentes no país, segundo algumas características importantes como sexo, idade, nível educacional, etc. Porém, mesmo no censo, nem todas as variáveis são obtidas entrevistando todas as pessoas. Devido aos altos custos envolvidos, e o uso das informações de forma mais agregada, outras características como renda, ocupação, etc., são obtidas através de amostras, entrevistando-se apenas os moradores de parte dos domicílios, algo em torno de um em cada dez domicílios. Outro exemplo de levantamento amostral bastante divulgado ultimamente são as pesquisas de intenção de votos.

Tipos de levantamento como os divulgados acima são mais "passivos," pois procuram identificar características da população sem interferir nos resultados. São as chamadas pesquisas de levantamento de dados (*survey*, em inglês). Outras vezes, deseja-se saber o que acontece com determinada variável quando as unidades são submetidas a tratamentos especiais controlados. Por exemplo, o uso de determinada vacina diminui a incidência de certa doença? A altura com que um produto é exposto na gôndola aumenta a oportunidade de venda? Nesses casos, é necessário trabalhar com grupos que recebam o tratamento e outros que sirvam como controle. São os conhecidos planejamentos de experimentos, ou simplesmente experimentação.

Outros critérios poderiam ser utilizados para identificar tipos de pesquisa. Na Figura 1.2, apresentam-se quatro possíveis critérios dicotômicos para classificar uma pesquisa. Só a combinação de suas alternativas já produziria 16 possíveis tipos de pesquisas quantitativas.

Neste livro, a preocupação maior será em apresentar pesquisas do tipo levantamento, com objetivos descritivos de dados simples obtidos de amostras. Eventualmente, serão tratados dados multivariados.

1.5.2 Métodos de coleta de dados

Escolhido o tipo de investigação, é necessário decidir que método será usado para obter os dados. Os comentários feitos a seguir serão muito mais adequados para

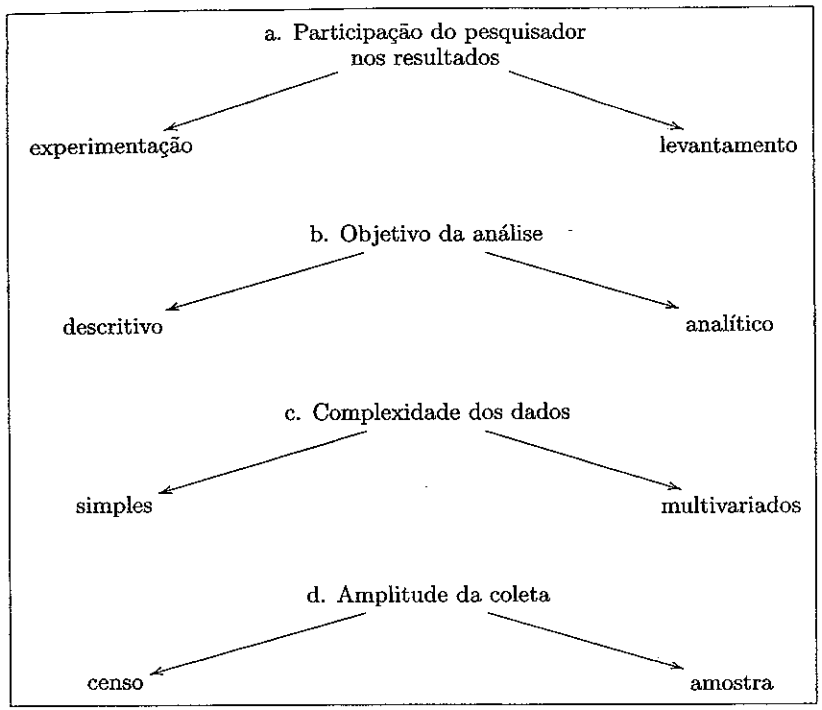


Figura 1.2: Critérios para classificar pesquisas

pesquisas amostrais, embora se apliquem também para outras situações.

Ter uma visão abrangente dos possíveis métodos de mensuração é muito útil para decidir qual seria o mais adequado para o levantamento que se pretende fazer. Um primeiro critério de classificação dos métodos pode ser aquele que avalia o processo de mensuração, ou seja, a utilização ou não de um instrumento formalizado para coleta das informações. Quando não utiliza instrumentos formalizados, o pesquisador vai anotando livremente o que observa em cada UE, procurando aprofundar aqueles aspectos que lhe pareçam mais interessantes. Assim, sempre são obtidas informações semelhantes que permitam a condensação em tabelas resumidas. São ilustrações deste método os chamados estudos de caso, de profundidade ou ainda de conteúdo. Por exemplo, para investigar como a popualação mais carente resolve seus problemas de saúde, pode-se começar perguntando a um líder comunitário como ele ajuda a resolver problemas de saúde apresentados por membros de sua comunidade. Em seguida, entrevistar um farmacêutico da região para saber qual o seu papel, depois a "benzedeira" local e assim por diante. Usualmente, este tipo de pesquisa

não é indicado para fazer inferências sobre a população, mas é muito útil para aprofundar o conhecimento sobre determinado assunto. Os instrumentos estruturados são mais usados em levantamentos e a sua versão mais conhecida é o questionário, preferencialmente com questões fechadas. Estes instrumentos formalizados traduzem a operacionalização dos conceitos que deverão ser obtidos, daí a importância de serem elaborados cuidadosamente, pré-testados e pré-analisados. Existe uma larga literatura no assunto, a qual é recomendada àqueles que pretendem fazer algum levantamento. Outros exemplos de instrumentos formalizados são: as planilhas de levantamento de estoques para medir consumo de certo produto; os "peoplemeters", pequenos aparelhos que registram o canal a que a televisão está ligada em pesquisas de audiência, e as cadernetas de consumo para o estabelecimento de um sistema de ponderação em pesquisa de custo de vida.

Um segundo critério para classificar os métodos de coleta dos dados é a forma de comunicação empregada: verbalizada ou não verbalizada. Estão classificados na segunda alternativa os chamados estudos observacionais. Na categoria verbalizada pode-se considerar a comunicação oral ou escrita. Estudos observacionais são usados, por exemplo, para analisar o comportamento de consumidores, para levantar opiniões em discussões de grupo, etc. Já a comunicação verbal é muito usada em levantamentos com populações humanas. A combinação destes critérios, aliados a outros, produzem uma gama de diferentes métodos de coleta espalhados pela literatura com os mais diversos nomes. Em amostragem, a combinação mais usada é a de comunicação verbal com mensuração estruturada. O uso de questionário com entrevista pessoal oral talvez seja a combinação mais utilizada em levantamentos. Variações muito comuns são as entrevistas pelo correio ou telefone.

Não há necessidade de ressaltar a importância do conhecimento do método de coleta dos dados no planejamento da amostragem. O número de elementos de um levantamento por correio costuma ser bem maior do que um semelhante, mas realizado com entrevista pessoal. Por quê?

1.5.3 Planejamento e seleção da amostra

Suponha que, após cuidadosa análise dos objetivos e orçamento, conclui-se que uma amostra é o procedimento indicado para análise de dados. **Amostra**, como o próprio nome indica, é qualquer parte da população.

Portanto, supõem-se já fixadas as unidades de análise, os instrumentos de coletas de dados, bem como a relação das unidades componentes da população, ou

seja, o sistema de referências. Desse modo, consideram-se também identificados e listados os elementos pertencentes à população de referência.

O propósito da amostra é o de fornecer informações que permitam descrever os parâmetros do universo, da maneira mais adequada possível. *A boa amostra permite a generalização de seus resultados dentro de limites aceitáveis de dúvidas.* Além disso, os seus custos de planejamento e execução devem ter sido minimizados. Embora estes conceitos sejam de fácil aceitação, a sua implementação não é assim tão trivial.

Qualquer amostra fornece informações, porém não é qualquer uma que permite estender os resultados para a população da qual foi retirada. Ouve-se freqüentemente o argumento de que uma boa amostra é aquela que é **representativa**. Quando se indaga sobre a definição de uma amostra representativa, a resposta mais comum é algo como: "Aquela que é uma microrrepresentação do universo". Mas para se ter certeza de que uma amostra seja uma microrrepresentação do universo para uma dada característica de interesse, deve-se conhecer o comportamento dessa mesma característica da população. Então, o conhecimento da população seria tão grande que se tonaria desnecessária a coleta da amostra.

Outras vezes, o significado da microrrepresentação confunde-se com o de uma amostra estratificada proporcional. Ou seja, a população é dividida em subpopulações (estratos) segundo alguma variável auxiliar, e de cada estrato sorteia-se uma amostra de tamanho proporcional ao seu tamanho. Este tipo de amostra não conduz obrigatoriamente a resultados mais precisos. Veja um exemplo a seguir.

Suponha que o objetivo é estudar a renda familiar de certa cidade. O conhecimento da geografia da cidade possibilita agrupar, aproximadamente, os bairros em mais ricos (**A**), médios (**B**) e pobres (**C**). Uma consulta aos registros da prefeitura permite afirmar que 10% dos domicílios pertencem à classe **A**, 30% à classe **B** e os restantes 60% à classe **C**. Se o orçamento garante entrevistar 1.000 domicílios, a amostra "representativa" seria selecionar 100 do estrato **A**, 300 do estrato B e 600 do estrato C. Observe que uma outra amostra "não representativa" que alocasse 600 ao estrato **A**, 300 ao **B** e 100 ao **C** pode apresentar resultados mais confiáveis. Basta lembrar que no estrato **C** os salários são muito parecidos, assim uma amostra de 600 domicílios seria um exagero. Já 100 unidades para o estrato **A**, onde as rendas variam muito, pode ser considerada muito pequena. Volte a contemplar este exemplo após estudar amostragem estratificada no Capítulo 4.

Diante da dificuldade em definir amostra representativa, os estatísticos preferem trabalhar com o conceito de amostra probabilística, que são os procedimentos

onde cada possível amostra tem uma probabilidade conhecida, a priori, de ocorrer. Desse modo, tem-se toda a teoria de probabilidade e inferência estatística para dar suporte às conclusões. Para generalizar as conclusões por meio de um outro procedimento, amostras intencionais, por exemplo, você deveria basear-se em teoria apropriada, digamos, teoria da intencionalidade, caso exista.

Embora este livro seja dedicado a estudar procedimentos da amostragem probabilística, na seção seguinte mencionam-se brevemente alguns outros tipos de procedimentos amostrais.

1.5.4 Tipos básicos de amostras

Jessen (1978) propõe um modelo interessante para identificar tipos de amostras, usando o cruzamento de dois critérios. O primeiro indica a presença ou ausência de um mecanismo probabilístico no plano de seleção da amostra, enquanto o segundo indica a existência ou não de um procedimento objetivo por parte do "amostrista" na seleção operacional da amostra. Procedimento objetivo é qualquer um, cujo protocolo descritivo é inequívoco. Ou seja, quando utilizado por pessoas distintas, produz a mesma amostra, ou uma com as mesmas propriedades. Um procedimento subjetivo é aquele que permite ao usuário usar seus julgamentos ou sentimentos para selecionar uma "boa" amostra. A combinação desses dois critérios permite criar os quatro tipos de planos amostrais apresentados na Tabela 1.1.

Tabela 1.1: Tipos de amostras

Critério do	Procedimento de seleção	
"amostrista"	probabilístico	não probabilístico
objetivo	amostras probabilísticas	amostras criteriosas
subjetivo	amostras quase-aleatórias	amostras intencionais

Neste livro, às vezes será usado imprecisamente o termo amostras como sinônimo de planos amostrais. Assim, por exemplo, pode aparecer mencionado tanto plano aleatório simples como amostras aleatórias simples para descrever um determinado procedimento de seleção. Entendem-se por amostras aleatórias simples as amostras obtidas através de um protocolo de seleção chamado plano aleatório simples.

Alguns exemplos de planos amostrais:

- probabilístico: amostragem aleatória estratificada proporcional;
- quase-aleatório: amostragem por quotas;

- criterioso: uso do conceito de cidade típica;
- intencional: júri de especialistas, voluntários.

1.5.5 Classificação de amostras probabilísticas

A qualidade do sistema de referências e outras informações disponíveis orientam o desenho do plano amostral mais adequado para atingir os objetivos da pesquisa. As múltiplas possibilidades dessas características podem gerar uma grande variedade de planos amostrais. Como sempre, a apresentação sistemática destas possibilidades fica mais fácil quando agrupadas por alguns critérios, gerando tipologias de planos amostrais. Usar-se-ão aqui os critérios propostos por Kish (1965) e resumidos na Figura 1.3.

A combinação dos resultados de cada um desses critérios apontados gera 32 possíveis planos amostrais. Por exemplo, usando-se as primeiras opções de cada critério tem-se o conhecido plano de **Amostragem Aleatória Simples**. Ou seja, cada unidade elementar é sorteada com igual probabilidade, individualmente, sem estratificação, e com um único estágio e seleção aleatória. Neste livro, serão abordados alguns destes planos e fornecidos instrumentos para que sejam exploradas as principais propriedades dos demais.

Quando o sistema de referências (SR) é *perfeito*, isto é, quando ele lista uma a uma todas as unidades de análise, é possível então usar um processo, onde cada unidade é sorteada diretamente com igual probabilidade de pertencer à amostra. A melhor maneira para definir este plano é descrevendo o processo de sorteio que seria o seguinte: "Da relação de unidades do SR sorteie, com igual probabilidade de pertencer à amostra, o primeiro elemento da amostra, repita o processo para o segundo e, assim sucessivamente, até sortear o último elemento programado para a amostra". As amostras assim obtidas definem o plano de **Amostragem Aleatória Simples** (AAS). Introduzindo-se o critério da reposição ou não da unidade sorteada antes do sorteio seguinte, obtém-se uma primeira dicotomia deste plano: Amostragem Aleatória Simples com e sem reposição (AASc e AASs). Do ponto de vista prático, dever-se-ia usar sempre amostras sem reposição, pois não estaria sendo incorporada nova informação se uma mesma unidade fosse sorteada novamente. Entretanto, do ponto de vista estatístico, a reposição recompõe o universo tornando mais fácil deduzir as propriedades dos modelos teóricos (independência). O plano AAS é o mais simples deles e serve como base para muitos outros. Além disso o plano AASc é aquele usualmente utilizado nos livros de inferência estatística.

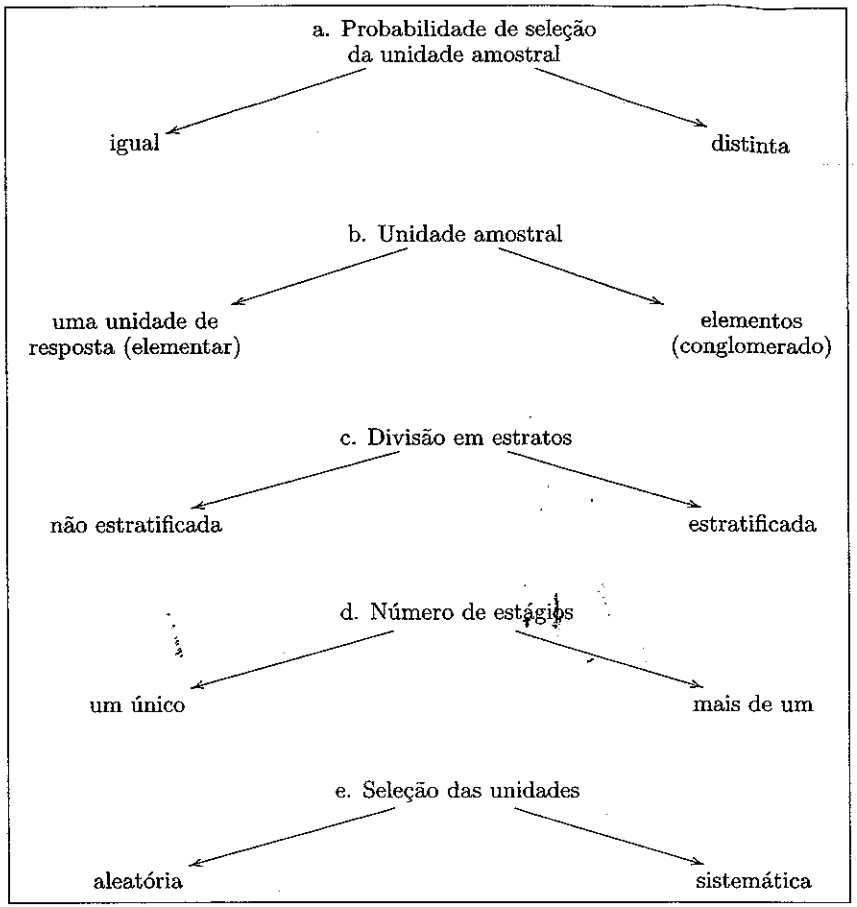


Figura 1.3: Critérios para classificar amostras probabilísticas

O sorteio das unidades com igual probabilidade é apenas uma estratégia que simplifica muito o desenvolvimento das propriedades matemáticas associadas ao plano, mas em algumas situações é conveniente sortear as unidades com probabilidades desiguais. Nesta última situação, e se ainda não for feita reposição, os modelos de análise tornam-se bastante difíceis de serem derivados.

Nem sempre, tem-se à disposição um sistema de referência completo. É muito comum ter-se uma relação descrevendo um grupo de unidades elementares. Por exemplo, em pesquisa sobre intenção de votos, onde a unidade elementar é eleitor, é muito comum contar com o SR como sendo a relação de domicílios. Ou seja, a unidade de sorteio será formada por um grupo de eleitores. Nem sempre a **unidade**

elementar coincide com a **unidade amostral**. Tecnicamente, esse agrupamento de unidades elementares será designado por **conglomerado**. Os planos amostrais selecionando conglomerados de unidades elementares serão chamados de **Amostragem por Conglomerados**.

Mesmo usando-se amostragem por conglomerados, o interesse continua sendo a análise das unidades amostrais, e a obtenção de informação é feita nas unidades elementares. Voltando-se ao exemplo acima, embora tenha sido sorteado um domicílio, deve-se obter a intenção de voto de cada eleitor do domicílio. Pode-se alegar, entretanto, que entrevistar todos os elementos do conglomerado é um desperdício, já que as opiniões no seu interior tendem a ser muito semelhantes. Isto sugere a adoção de um sorteio em dois estágios: na primeira etapa sorteia-se o conglomerado (domicílio) e, dentro do conglomerado selecionado, sorteia-se a unidade elementar (eleitor). São os chamados planos de amostragem em **múltiplos estágios**. Este é um tipo de amostragem muito usado em populações humanas, onde inicialmente se sorteiam as cidades, depois os bairros, quarteirões, domicílios e finalmente moradores. O uso de várias unidades de sorteio define em cada estágio uma diferente unidade amostral. Assim, no primeiro estágio, tem-se a **Unidade Primária de Amostragem (UPA)**, no segundo estágio a **Unidade Secundária de Amostragem (USA)**, etc.

O uso de informações adicionais é fundamental para aprimorar um desenho amostral. Por exemplo, em uma pesquisa sobre renda familiar média, conhecem-se de antemão as regiões da cidade onde predominam moradias de diferentes classes de renda. Esse conhecimento pode ser usado para definir subpopulações homogêneas segundo a renda, e então sortear amostras dentro de cada uma das regiões. Este procedimento é conhecido como a divisão da população em estratos e, conseqüentemente, definem os **Planos de Amostragem Estratificada**. A estratificação procura explorar a idéia de que, quanto mais homogênea for a população, mais preciso são os resultados amostrais. Suponha por absurdo que um processo de estratificação consiga reunir em um estrato todas as famílias com uma mesma renda. Para estimar este valor basta então sortear uma única família desse estrato. Quase todos os planos amostrais reais adotam a estratificação em algumas de suas etapas. A maneira de alocar as unidades amostrais pelos estratos define diferentes famílias de Amostragem Estratificada que serão estudadas nos capítulos correspondentes.

Finalmente, o sorteio das amostras pode ser feito aleatoriamente um a um, ou então criar conglomerados especiais agrupando unidades eqüidistantes umas das outras e sorteando um ou mais destes conglomerados. Por exemplo, pode-se formar

um conglomerado contendo as unidades elementares com as posições 1, 11, 21, 31, etc. do SR; outro conglomerado contendo os elementos 2, 12, 22, 32, etc. e assim por diante. Desse modo, haveria 10 possíveis conglomerados artificiais e o sorteio de um deles forneceria uma amostra de 10% do total da população. Esse procedimento, muito usado no passado, é conhecido como sorteio sistemático. Ele facilita muito o sorteio das unidades, mas introduz alguns problemas técnicos difíceis de serem resolvidos.

1.5.6 Estimadores e erros amostrais

Suponha que, a esta altura da pesquisa, já estão definidos e escolhidos: o sistema de referências, a(s) variável(eis) e respectivo(s) parâmetro(s) de interesse, o plano amostral e tamanho de amostra; resta então escolher a(s) característica(s) da amostra que será(ão) usada(s) para responder aos objetivos específicos da pesquisa. Para facilitar a exposição, suponha que o interesse principal é conhecer um parâmetro θ associado a uma variável Y de interesse da população. A questão passa a ser que estatística (característica) t será usada para estimar θ. A teoria para escolha do "melhor" estimador encontra-se desenvolvida nos livros de Inferência Estatística e os próximos capítulos serão dedicados a estudar algumas propriedades de estimadores simples para alguns planos amostrais particulares. Nesta seção, dar-se-á um tratamento menos formal para o assunto.

O uso de um levantamento amostral introduz algum tipo de erro, que pode ser resumido na diferença entre o valor observado na amostra e o parâmetro de interesse na população. Esta diferença pode ocorrer apenas, devido à particular amostra escolhida, ou então, devido a fatores externos do plano amostral. O primeiro são os chamados erros amostrais, objeto de avaliação estatística do plano amostral. Em seção futura, serão estudados alguns outros tipos de erros envolvidos em um levantamento amostral. Evidentemente, a avaliação de um plano amostral passa pelo conhecimento e mensuração da magnitude possível do erro global, ou seja, aquele englobando os dois tipos de erros.

O estudo do erro amostral consiste, basicamente, em estudar o comportamento da diferença $t - \theta$, quando t percorre todas as possíveis amostras que poderiam ser formadas através do plano amostral escolhido. Se o valor esperado desta diferença for igual a zero, tem-se um estimador **não viesado**. Já o valor esperado do quadrado desta diferença, o **erro quadrático médio (EQM)** informa sobre a precisão do estimador. Procuram-se usualmente estimadores com baixos EQM. Quando o

estimador é não viesado, o EQM passa a ser a variância do estimador, calculada em relação à distribuição amostral do estimador. Para recuperar a mesma unidade da variável, usa-se o desvio padrão, que nada mais é que a raiz quadrada da variância. Neste caso particular, o desvio padrão recebe o nome de erro padrão do estimador, que pode ser visto como indicador do erro médio esperado pelo uso deste estimador e deste plano amostral.

Do ponto de vista estatístico, o objetivo ao escolher-se um estimador e desenhar um plano amostral, é poder controlar o erro padrão usualmente traduzido pelos intervalos de confiança que podem ser construídos. Mais ainda, o objetivo é conseguir erro padrão baixo.

O uso de informações adicionais para melhorar as estimativas, como no caso da estratificação acima, é muito empregado em amostragem. Entretanto, essa informação às vezes é usada para melhorar os estimadores, e não o plano amostral. Por exemplo, deseja-se estimar através de amostragem o número de desempregados em determinada região. Os dados do registro civil fornecem informações precisas sobre a população em idade ativa (PIA - pessoas com mais de 15 anos). Pode-se usar a taxa de desemprego em relação à PIA obtida na amostra, combinada com os dados do registro civil para produzir melhores estimativas. Neste livro, serão analisados dois tipos de estimadores que incorporam informações adicionais através de variáveis auxiliares: razão e regressão.

1.5.7 Tamanho da amostra

O erro padrão do estimador, como será visto em capítulos posteriores, decresce à medida que aumenta o tamanho da amostra. Assim, um ponto-chave de um levantamento amostral é a fixação do tamanho da amostra.

Uma amostra muito grande pode implicar em custos desnecessários, enquanto uma amostra pequena pode tornar a pesquisa inconclusiva. Suponha um levantamento amostral, cujo objetivo é prever qual dentre os dois únicos possíveis partidos terá maior porcentagem de votos válidos - excluídos nulos e brancos. Aceite também que foi utilizado um plano amostral aleatório simples (AAS) e um dos partidos obteve 56% dos votos. Caso tivesse sido usada uma amostra de 100 eleitores, o intervalo de 95% de confiança indicaria um número entre 46% e 66%, portanto inconclusivo para afirmar se o partido ganharia ou não a eleição. Já uma amostra de 400 eleitores indicaria o intervalo entre 51% e 61%, sugerindo a vitória do partido. Por outro lado, uma amostra de 1.600 eleitores definiria o intervalo entre 53,5% e 59,5%,

implicando no uso desnecessário de 1.200 unidades a mais. O problema real é muito mais complexo que o apresentado aqui, mas o exemplo dá uma boa ilustração dos problemas estatísticos envolvidos na determinação do tamanho da amostra.

Um dos aspectos pouco discutidos em cursos de amostragem é aquele associado aos custos de um levantamento. Este tópico é fundamental para o delineamento de toda a pesquisa, desde a definição dos objetivos possíveis de serem respondidos, passando pelo tamanho da amostra economicamente viável e chegando até a escolha da sofisticação do modelo de análise a ser adotado. Recomenda-se àqueles que venham a se dedicar à prática de amostragem que estudem mais profundamente este aspecto, podendo consultar principalmente o livro de Kish (1965) e Lansing e Morgan (1971).

Como já foi mencionado, muitas vezes a precisão estatística desejada para a pesquisa esbarra nas limitações impostas pelo orçamento, obrigando a decidir entre realizar a pesquisa baixando a precisão desejada ou não realizar o levantamento. Isto nos remete ao compromisso para fixar o tamanho da amostra, ou mesmo para a pesquisa como um todo, em *procurar dentro das restrições impostas pelo orçamento, desenhar uma amostra que atinja os objetivos, produzindo estimativas com a menor imprecisão possível.*

Embora neste livro a determinação do tamanho da amostra será sempre feita levando em conta os aspectos da precisão estatística, acredita-se que, na maioria dos casos, a decisão segue a proposição acima. Isto é, as limitações orçamentárias definem o tamanho da amostra e então estima-se a precisão possível. Se os dois interesses coincidirem, então se realiza a pesquisa.

1.5.8 Censo ou amostragem

Usa-se aqui o termo levantamento tanto para indicar a pesquisa feita para um recenseamento (ou censo), como para uma amostra. O que as diferencia é o número de unidades entrevistadas: no primeiro são todas e no segundo uma parte.

Muitas pessoas acreditam que apenas através do censo é que se pode conhecer a "verdade" sobre a população. É claro que, em igualdade de condições, o censo produz resultados mais precisos que a amostra. Entretanto, como já foi mencionado, limitações orçamentárias impõem restrições que podem tornar o levantamento amostral mais fidedigno do que o censo. Imagine uma pesquisa com orçamento fixo, para conhecer o estado de saúde da população. Pode-se fazer um censo usando questionário como instrumento de coleta de informação, ou então uma amostra com

exames clínicos e laboratoriais feitos por médicos e paramédicos. Parece que a segunda opção produzirá resultados muito mais informativos e precisos que o primeiro.

Recomenda-se o uso de censo quando a população é pequena, quando há erros amostrais grandes, informações baratas ou alto custo em tomar decisões erradas. O bom senso deve prevalecer em algumas decisões. Por exemplo, se a precisão estatística sugere uma amostra maior do que a metade da população é bem mais razoável fazer um censo, desde que os custos o permitam. O censo seria indicado para uma pesquisa sobre a participação dos chefes de departamentos em uma universidade, na definição da política de recrutamento de novos docentes.

Em contraposição, deve-se usar amostragem quando a população é muito grande e/ou o custo (em dinheiro e tempo) de obter informações é alto. Seria recomendada se, na universidade do exemplo acima, se quisesse conhecer a opinião dos alunos sobre a qualidade dos professores em sala de aula.

1.6 Coleta dos dados (trabalho de campo)

Para o sucesso de um levantamento, não basta um plano amostral tecnicamente perfeito, se as informações não forem recolhidas com fidedignidade. Imagine uma pesquisa sobre salários, onde o entrevistador não foi instruído para anotar se a informação refere-se a salário líquido ou bruto. Como será possível analisar os dados? Ou ainda, em pesquisa domiciliar onde apenas um elemento da casa será entrevistado, deixar esta escolha para o entrevistador. Sem dúvida, ele escolherá um membro presente na casa, na hora da entrevista, introduzindo um viés na pesquisa. Provavelmente, este levantamento terá uma proporção bem maior de mulheres. Se não forem tomados cuidados, o trabalho de campo pode arruinar totalmente uma pesquisa. Assim, deve-se planejar e usar procedimentos que minimizem os erros, ou viéses introduzidos na coleta de dados.

Jessen (1978) resume estes cuidados na seguinte frase: *"As medidas são aquelas óbvias; selecionar boas pessoas, treiná-las bem e verificar se fazem o trabalho corretamente"*.

O volume de trabalho para operacionalizar essas medidas irá depender principalmente do tamanho da pesquisa e do fato de a pesquisa ser pontual (ad-hoc) ou periódica. Para pesquisas pequenas, o treinamento de pessoal envolvido é bem reduzido, podendo chegar ao caso de ser apenas o próprio pesquisador. Em pesquisas periódicas o esforço deve ser maior para elaborar manuais e material de consulta que serão usados freqüentemente. Entretanto, pode-se apresentar sucintamente alguns

comentários em como evitar viéses nos cuidados mencionados por Jessen.

Recrutamento. Para pesquisas grandes, realizadas uma única vez, recomenda-se a contratação de empresas especializadas que possuam pesquisadores profissionais e que estejam acostumados com a aplicação e administração deste tipo de trabalho. A alternativa, freqüentemente mais barata, será a de executar o trabalho todo de contratar entrevistadores, listadores, supervisores, checadores, etc., cada um deles com um perfil próprio, desenvolver programas de qualidade da coleta, etc. Com uma seleção imprópria ou "caseira", corre-se o risco de pagar caro pelo noviciado. Para pesquisas periódicas, e com a necessidade constante de renovação e substituição de pessoas envolvidas, pode-se criar um núcleo permanente de seleção de pessoal, com a vantagem adicional de a escolha ser dirigida para os objetivos específicos do trabalho.

Treinamento. O pessoal de pesquisa deve ser bem treinado não apenas com os conceitos, definições, uso do instrumento de mensuração, etc., mas também com os melhores procedimentos para extrair as informações desejadas. Existem técnicas bem desenvolvidas acerca de como abordar as pessoas, de postura, de entonação de voz e outras. Ou ainda, o treinamento para uma pesquisa frente a frente é bem diferente de uma por telefone. Em pesquisas muito grandes, os problemas envolvidos com o treinamento são enormes e requerem muitas vezes o uso de mecanismos bastante especiais. Apenas imagine os cuidados que devem ser tomados para o treinamento de mais de 150 mil entrevistadores para a realização do censo populacional brasileiro. Nestes casos, e na maioria deles, recomenda-se a adoção de manuais escritos para cada uma das tarefas: listagem, entrevistas, checagem, codificações, etc.

Embora o treinamento procure prever todas as situações que serão encontradas, é preciso dar instruções sobre situações imprevistas. Por exemplo, na casa sorteada, há mais de um domicílio e várias famílias, ou ainda, não se consegue encaixar a profissão do chefe em nenhum dos casos listados. O entrevistador deveria entrar em contato com a supervisão, ou então anotar o maior número possível de informações para possível correção no escritório.

Verificação. É importante que se tenha um processo de controle contínuo da qualidade do trabalho de campo. A verificação deve ser realizada em várias etapas do trabalho do pesquisador. No início da pesquisa, deve-se fazer um acompanhamento mais meticuloso para verificação do entendimento correto dos

conceitos, da identificação exata das unidades selecionadas e de resposta, aprimorando e corrigindo-as imediatamente. Além de verificações rotineiras, deve-se ter um plano de verificação aleatória, onde uma subamostra é reentrevistada para apurar desde fraude até a qualidade das informações obtidas. Este procedimento permite avaliar a magnitude de alguns viéses introduzidos pelo trabalho de coleta de dados.

A supervisão de campo deve estar em permanente contato com os responsáveis do planejamento para obter os esclarecimentos sobre questões ambíguas e decisões a serem tomadas para casos imprevistos. Também, o contato com os responsáveis pelo processamento dos dados ajuda a esclarecer e remover informações desencontradas e os erros mais comuns cometidos pelo pessoal de campo.

Registro. Muitas ocorrências e decisões imprevistas acontecem nesta fase e é muito importante que se mantenha um registro atualizado das mesmas para futuras avaliações do desempenho do levantamento. As estatísticas e qualificações sobre as unidades perdidas e as incluídas indevidamente é que permitirão a descrição pormenorizada da *população amostrada*. As dúvidas e inadequações apresentadas pelos entrevistadores, bem como os esclarecimentos prestados ajudarão a entender a qualidade, significado e fidedignidade das respostas obtidas.

1.7 Preparação dos dados

Se não for devidamente avaliada, planejada e executada, a construção inicial do banco de dados pode-se tornar a etapa mais demorada de um processo de levantamento de informações.

Usando-se uma imagem bastante simplificada, pode-se descrever o banco de dados como sendo uma matriz de $n+1$ linhas por $p+1$ colunas. As linhas correspondem às n unidades respondentes e as colunas, às p variáveis de interesse. A primeira coluna descreve a identificação da unidade respondente, enquanto a primeira linha denomina as variáveis. A célula (i,j) contém os dados codificados da j-ésima variável para a i-ésima unidade respondente. Estes dados devem estar disponíveis em um meio que permita o fácil acesso e manipulação. Imagina-se um meio eletrônico conveniente.

A construção desta tabela exige: (i) transcrição; (ii) minucioso escrutínio da qualidade e (iii) disponibilização das informações.

Transcrição. Esta tem sido a fase mais demorada do processo, porém tem sido aquele segmento onde a tecnologia vem apresentando soluções bem competentes. Quanto menos haja intervenção na transcrição de um meio para outro, menor a possibilidade de introdução de erros na pesquisa. Deve-se procurar balancear o custo de uso de recursos mais sofisticados com a qualidade e rapidez para a execução desta tarefa.

Qualidade dos dados. Antes de liberar os dados para a análise, deve-se ter certeza da boa qualidade dos mesmos. O escrutínio crítico dos dados passa pela identificação de erros de transcrição, de inconsistências e outros tipos de enganos. A correção pode ser feita com a ajuda da lembrança e interpretação dos pesquisadores, com o apoio de processos automáticos e, quando for necessário, revisitando a unidade sorteada.

A utilização de programas automáticos de análise da consistência lógica das respostas é uma das ferramentas mais poderosas na detecção de vários tipos de erros. O conhecimento substantivo do instrumento de pesquisa associado à habilidade do pesquisador possibilita a construção de bons mecanismos de detecção automática de erros. Hoje em dia, com o uso de instrumentos eletrônicos de entrada de dados, este tipo de controle vem sendo feito no ato de coleta, não aceitando a entrada de dados inconsistentes.

Ainda nesta fase, quando programado, é necessária a utilização de procedimentos de imputação de dados. São usados principalmente para imputar valores baixos deixados em branco para itens fundamentais do levantamento, ou ainda para substituir dados incompatíveis. Como exemplo desta última situação, temos procedimentos especiais para transformar dados sobre salários líquidos em brutos.

Em grandes pesquisas, o treinamento da equipe de transcrição e crítica deve seguir os mesmos cuidados apresentados na coleta. Manuais de críticas garantem a homogeneidade dos critérios empregados nas correções e imputações.

Banco de dados. Terminadas a entrada e a crítica das informações coletadas, a base de dados está quase pronta e apta a receber os primeiros tratamentos estatísticos. Para completá-la e facilitar o sucesso, é muito importante que

esta base venha acompanhada de informações precisas sobre o seu conteúdo. É comum encontrar no banco de dados apenas uma coleção de algarismos e símbolos, sem nenhuma descrição do significado das variáveis, sua formatação, recomendações sobre a qualidade, sistema de ponderação, etc. Desse modo, o banco de dados deve vir acompanhado de documentação que permita a qualquer pessoa, vinculada ou não à pesquisa, usar os dados sem muita dificuldade. Voltaremos a tocar nesse assunto na Seção 1.11.

1.8 Análises estatísticas

A partir da base de dados, várias análises podem ser feitas, cada uma delas com seu objetivo específico.

Análise exploratória. Na ausência de uma expressão melhor, considerar-se-á este nome para indicar as primeiras manipulações estatísticas. Deve-se começar estudando a distribuição de freqüências de cada variável (ou campo) do banco de dados, acompanhada de algumas medidas e resumos. Além de tornar o pesquisador mais íntimo dos dados, a análise exploratória permite-lhe identificar erros não detectados pela crítica, a existência de elementos desajustados, quantidade de respostas em branco e, com um pouco mais de sofisticação, a descoberta de possíveis viéses introduzidos pelos entrevistadores ou outro trabalho de campo. É muito comum encontrar determinadas características com alta concentração de respostas em um nível de categoria, tornando praticamente inútil o uso desta "variável" nos estudos. O emprego de tabelas cruzadas para algumas características decompostas pelos estratos, ou por fatores geográficos, econômicos, demográficos. etc., permite adquirir maior conhecimento de seus significados. A comparação com resultados de outras pesquisas confiáveis, tais como os censos, permite avaliar a qualidade do levantamento.

Plano tabular. Com esse título, entende-se aquele conjunto mínimo de tabelas e modelos estatísticos que foram definidos "a priori" para responder aos objetivos iniciais da pesquisa.

O exercício, realizado antes da obtenção dos dados, de imaginar operacionalmente como os dados recolhidos na pesquisa responderiam aos objetivos da pesquisa, além de ajudar, e muito, o planejamento amostral, evita divulgar os resultados em prazos distantes do trabalho de campo tornando-os desinteressantes. Serve também para que sejam previamente preparados, escolhidos e testados os programas computacionais necessários para sua execução. Usualmente, estas primeiras respostas são fornecidas por tabelas de duplas entradas, daí o nome de plano tabular.

Junto com a divulgação da aplicação do plano tabular, recomenda-se que também sejam apresentados os erros amostrais, permitindo avaliar qual a confiabilidade apresentada pela pesquisa. Para pesquisas com um número muito grande de variáveis, deve-se procurar modos adequados e resumidos para divulgação dos erros. Pode-se encontrar exemplos de como divulgar os erros amostrais, consultando-se os compêndios de metodologia publicados pelo IBGE.

Análises adicionais. Os levantamentos estatísticos de um modo geral possuem muito mais informações do que aquelas usadas para responder aos objetivos iniciais. Pode-se, em uma segunda etapa, voltar a explorar os dados para testar novas hipóteses ou mesmo para especular sobre relações inesperadas. Um único levantamento amostral sobre condições de vida realizado pela Fundação SEADE produziu mais de 10 trabalhos em um período de 3 anos. Durante pelo menos 10 anos, até que um novo seja realizado, os censos demográficos são investigados, em várias dimensões e por pesquisadores de diversas instituições.

Também os modelos de análise podem ser bem mais sofisticados do que simples tabelas descritivas, desde que haja tempo para investigar a adequação e pertinência dos mesmos. Na mencionada pesquisa da Fundação SEADE, alguns estudos foram novamente analisados, empregando-se modelos para dados categóricos e outros modelos multivariados.

Uma das consequências mais importantes da análise dos dados é a possibilidade de criação de novas variáveis (índices) resultantes da combinação de outras, e que descrevam de maneira mais adequada os conceitos pretendidos. Voltando-se à pesquisa do SEADE, usaram-se combinações do grau de educação do chefe e de um segundo membro da família para criar um grau de educação da família. De modo mais sofisticado, e com técnicas estatísticas, criou-se uma condição de qualidade de emprego da família.

1.9 Erros

Todo levantamento, amostral ou não, está sujeito a produzir diferenças entre o parâmetro populacional θ, de interesse, e o valor t empregado para estimá-lo. A

diferença $t - \theta$ é considerada como o **erro da pesquisa**. Vários fatores podem agir sobre esta diferença e fazem parte da avaliação detectá-las, tentar medi-las e avaliar suas conseqüências. Para facilitar a exposição, dividir-se-ão os fatores que afetam esta diferença em dois grandes grupos:

- erros devidos ao plano amostral;
- erros devidos a outros fatores.

Os primeiros deles, já mencionados na Seção 1.5.6, talvez sejam equivocadamente chamados de erro. Melhor seria chamá-los de desvio, objeto controlado pelos processos estatísticos que serão devidamente tratados nos demais capítulos deste livro. Estes desvios tendem a desaparecer com o crescimento do tamanho da amostra.

Os erros do segundo grupo são resultantes de inadequações dos processos de mensuração, entrevistas, codificações, etc. Eles permanecem mesmo em censos populacionais. Eles serão analisados nas seções abaixo.

A qualidade do levantamento está associada à capacidade do pesquisador em evitar, ou se não for possível, procurar manter esta diferença em níveis aceitáveis. O conceito mais amplo da qualidade do levantamento deveria ser expresso em uma medida do erro total, contendo a mensuração dos erros amostrais e avaliações, qualitativas ou quantitativas, dos possíveis efeitos dos demais erros. Para estes últimos, é extremamente desejável que seja feita uma interpretação substantiva das possíveis conseqüências das direções e magnitudes dos seus vieses.

1.9.1 Erros amostrais

Conforme já definido anteriormente, considera-se um erro amostral aquele desvio devido apenas ao processo amostral, e não de problemas de mensuração e obtenção das informações.

Quando o plano adotado é do tipo probabilístico, a qualidade traduz-se pela estimativa do seu erro padrão, como já foi definido anteriormente. Boa parte deste livro dedicar-se-á ao estudo do desenvolvimento de técnicas para mensurar este erro. Entretanto, para alguns planos amostrais bastante complexos o conhecimento estatístico existente não é suficiente para prover expressões explícitas para estes erros, sendo necessário o recurso de técnicas especiais aproximadas. Às vezes, por ignorância ou facilidade de cálculo, emprestam-se fórmulas de um plano mais simples para o cálculo do erro padrão de outros planos amostrais mais complexos, praticando-se um "erro técnico". Quando esta escolha é consciente, sugere-se que o

pesquisador informe este fato, acompanhado do possível tipo de distorção introduzida por esta decisão.

Já para planos não probabilísticos, o maior desafio, e de difícil aceitação, é o de estender o resultado da amostra para a população e o de prover uma teoria para mensurar o erro cometido. Esta avaliação é feita usualmente através do arrazoado qualitativo, nem sempre convincente.

1.9.2 Erros não amostrais

Quando o desvio ocorre devido a fatores independentes do plano amostral, e que ocorreriam mesmo se a população toda fosse investigada, será considerado como erro não amostral. Eles podem aparecer em qualquer etapa do levantamento amostral (definições, coleta de dados, codificações e análise), e se não forem identificados e avaliadas as possíveis distorções introduzidas, podem comprometer seriamente um plano amostral tecnicamente perfeito.

Um modo de analisar este tipo de erro é explicar os seguintes pontos:

i. a etapa onde o erro ocorreu;

ii. quais as causas possíveis;

iii. a correção empregada, caso haja;

iv. e a avaliação qualitativa e/ou quantitativa, dos efeitos sobre os resultados.

Alguns autores preferem agrupar os erros na seguinte classificação dicotômica:

a. erros de observação, ocorridos durante o levantamento dos dados;

b. outros erros, ocorridos em outros momentos.

Recomendamos ao leitor interessado buscar mais informações em livros como o de Jessen (1978).

Apresentam-se abaixo, de modo bem abreviado, algumas possíveis ocorrências de erros não amostrais.

A. Unidades perdidas (falta de resposta), fatores para não resposta:

   i. Falta de resposta total

      a. Falta de contato com a unidade

      b. Recusa

      c. Abandono durante a pesquisa

d. Incapacidade em responder

e. Perda de documento

ii. Falta de resposta parcial

a. Recusa em questões sensíveis - renda

b. Incompreensão

c. Dados incoerentes

B. Falhas na definição e administração:

a. Sistemas de referência

i. Erros de omissão (cobertura incompleta), exclusão de elementos de interesse. Resulta de diferenças entre as diversas populações.

ii. Erros de comissão. Inclusão de elementos não sorteados ou de outras populações.

b. Efeito do entrevistador

c. Insuficiência do questionário - redação

d. Erros de codificação e digitação

C. Avaliação das conseqüências:

a. Comparação com resultados de outras pesquisas

b. Efeito do processo de imputação, caso tenha sido usado

c. Programas de consistência de dados

d. Volume de não respondentes

e. Diferença de perfil de respondentes e não respondentes

1.10 Apresentação dos resultados

O relatório do plano amostral presta contas para uma determinada audiência sobre os procedimentos adotados para escolha e coleta das unidades elementares portadoras dos dados de interesse do levantamento.

Um plano amostral tecnicamente perfeito e corretamente aplicado pode não ter sua qualidade reconhecida, devido a um relatório mal escrito e/ou mal organizado. As propostas para desenvolver competências em se comunicar são bem conhecidas e não serão abordadas aqui. Apenas insiste-se, que sejam consultadas as bibliotecas

especializadas e praticadas as recomendações sugeridas. Há muita similaridade entre relatórios descrevendo planos amostrais e outros tipos de relatórios científicos. Desse modo, sugerimos consultar também livros que tratam deste assunto, tais como Eco (1977) ou Babbie (1999). Ressaltam-se a seguir, na elaboração do relatório, alguns pontos específicos que devem ser considerados.

Como os relatórios podem ter diferentes formatos e tamanhos, deve-se em primeiro lugar decidir para qual audiência eles estão sendo escritos. Caso seja dirigido a um público afeito à linguagem de amostragem, será possível usar um vocabulário mais técnico do que aquele destinado ao público leigo.

Algumas vezes, o relatório do plano amostral é apenas uma pequena parte dentro da seção de metodologia, devendo então ser bastante conciso e direto. Outras vezes, ele é o produto final de seu trabalho, devendo incluir a descrição de todas as etapas, bem como a descrição, construção e análise do banco de dados e, neste caso o relatório será muito mais amplo e detalhado.

Sugere-se como prática de trabalho escrever sempre um relatório completo, elaborado conforme o desenrolar do levantamento. Ele servirá como uma espécie de diário e memória. A partir dele, você poderá extrair outros produtos que sejam de interesse. Você poderá usar os itens mencionados no Apêndice B como guia, sem a necessidade de respeitar a ordem apresentada.

Resumindo, qualquer que seja o tipo de relatório usado, ele deve mencionar pelo menos os seguintes itens: propósitos, as diversas populações. sistema de referência, unidades amostrais, plano de seleção, procedimento de coleta, desempenho da amostra, tamanho, sistema de ponderação, fórmulas para os erros amostrais e avaliações dos possíveis efeitos dos erros não amostrais.

Quando o relatório também inclui a análise, distinga bem os resultados descritivos da amostra dos que fazem inferências populacionais. Para grandes volumes de dados, onde a apresentação dos erros amostrais pode poluir e dificultar a leitura de cada tabela, sugere-se a adoção de procedimentos agregados que avaliem erros aproximados globais. Grandes institutos de pesquisa costumam usar este tipo de apresentação para os erros amostrais (consulte, por exemplo, as publicações do IBGE).

1.11 Divulgação do banco de dados (disponibilidade)

Falta à maioria dos bancos de dados, obtidos por levantamentos amostrais, uma documentação bem elaborada "que descreva a utilidade das variáveis e liste os vínculos

entre os códigos e os atributos que compõem as variáveis" (Babie, 1999), conforme mencionado na Seção 1.7. Essa ausência deve-se ao fato de que, na maioria das vezes, os dados serão produzidos e analisados por uma única pessoa ou grupo, tornando-se aparentemente dispensável esse trabalho. Entretanto, esse descuido já causou muitos prejuízos, tempo perdido e duplicação de trabalho, ao se analisar o mesmo banco de dados em ocasiões distintas.

Manter um banco de dados organizado e documentado deve ser uma preocupação prioritária dos "amostristas" e dos analistas de dados. Os primeiros usam-no para bem caracterizar os sistemas de ponderação e recodificções, e os segundos para descrever as recodificações, novas variáveis e indicadores criados.

O Banco de Dados junto com esse dicionário descritivo permite oferecer mais um serviço: disponibilizar a pesquisa para um público maior, graças as facilidades oferecidas hoje pela comunicação eletrônica. Como orientação para organizar esse serviço, sugere-se consultar os bancos de dados disponíveis no IBGE e SEADE.

Exercícios

1.1 Apresente uma questão ligada à sua área de interesse e que poderia ser respondida por um levantamento amostral. Aproveite para definir claramente quais seriam os seguintes conceitos na sua pesquisa:

a. unidade de pesquisa;

b. população;

c. instrumento de coleta de dados;

d. unidade respondente;

e. possível sistema de referência;

f. unidade amostral mais provável;

g. unidades amostrais alternativas.

Discuta também como você fixaria o tamanho da amostra a outros tópicos que achar relevantes.

1.2 Desenhe um plano amostral, ressaltando os pontos discutidos neste capítulo para responder ao seguinte problema: "*Deseja-se conhecer o número total de palavras existentes no livro texto Elementos de Amostragem por Bolfarine e Bussab*".

1.3 Planeja-se uma pesquisa para determinar a proporção de crianças do sexo masculino com idade inferior a 15 anos, moradoras de uma cidade. Sugerem-se três procedimentos:

a. Para cada menino de uma amostra de n meninos (retirada da população de meninos menores de 15 anos) pede-se informar quantos irmãos e irmãs ele tem;

b. Toma-se uma amostra de n famílias e pergunta-se o número de meninos e meninas menores de 15 anos existentes;

c. Procura-se casualmente n crianças de 15 anos e, além de anotar o sexo do entrevistado, pergunta-se o número de irmãos e irmãs que eles possuem na faixa etária de interesse.

Analise os planos amostrais acima e justifique suas afirmações. Diga e justifique qual deles você usaria, ou então proponha um outro.

1.4 A comissão de pós-graduação de sua universidade pretende fazer uma pesquisa cuja, população-alvo é formada por todos os alunos de pós-graduação. Um dos principais objetivos é estimar a proporção dos favoráveis a uma determinada mudança nas exigências do exame de qualificação, e espera-se que essa proporção seja da ordem de 5%. Imagine a situação na sua universidade e proponha um plano amostral, destacando: sistema de referência, tamanho da amostra, UPA, USA, fórmulas de estimadores e variâncias.

1.5 Sugira um esquema amostral aproximado para escolher amostras aleatórias nos seguintes casos:

a. Árvores em uma floresta;

b. Crianças abaixo de 5 anos e que tiveram sarampo;

c. Operários em indústrias têxteis.

Em cada caso, sugira uma variável que poderia ser estudada, qual a lista de elementos a que você teria acesso e faça as suposições (razoáveis) necessárias para resolver o problema.

1.6 Uma rede bancária tem filiais espalhadas por todo o país e seu pessoal especializado (cerca de 20 mil) é removido freqüentemente de um ponto para

outro. Deseja-se selecionar uma amostra de 10% do atual pessoal especializado, para uma pesquisa contínua durante os próximos anos. Pretende-se obter informações sobre o progresso da firma, mudança de emprego, etc. A seleção de uma amostra aleatória de 2 mil indivíduos seria muito cara, por questões de identificação. Foi proposto então que se sorteasse uma letra (digamos S) e todos os funcionários com sobrenomes começando com essa letra fariam parte da amostra. A inicial do sobrenome tem a vantagem de ser facilmente identificável, porque as fichas dos funcionários são arquivadas em ordem alfabética. Quais as críticas que você faria a este plano? Sugira um plano "melhor", mas ainda baseado nas vantagens da ordem alfabética. Descreva sucintamente o seu novo plano.

1.7 Descreva sucintamente como pode ser incorporado num plano amostral o conhecimento de variáveis auxiliares da população.

1.8 O IME-USP formou no ano passado a sua sétima turma de bacharéis em Estatística e deseja fazer um levantamento através de amostra, com múltiplos propósitos. Os principais objetivos são: estimar a proporção de formandos que realmente exercem a profissão e estimar o salário médio. Proponha um esquema amostral e aponte as dificuldades que provavelmente serão encontradas.

1.9 Faça uma lista de pontos essenciais para propor, executar e analisar um levantamento amostral.

1.10 Um pesquisador pretende estimar o consumo médio de água por domicílio em uma cidade. Discuta as vantagens e desvantagens em usar as seguintes UPAs:

a. Unidade domiciliar;

b. Blocos de domicílios: casa, prédio de apartamentos, vilas, etc.;

c. Quarteirões.

1.11 Um engenheiro florestal quer estimar o total de pinheiros de uma área reflorestada com diâmetro superior a 30 cm. Discuta como planejar uma pesquisa amostral para esse problema.

1.12 Um especialista em trânsito quer estimar a proporção de carros com pneus carecas na cidade de Pepira. Ele poderá usar sorteio de carros ou grupos de carros em estacionamento ou na rua. Discuta as vantagens de um ou de outro procedimento. Qual você usaria?

1.13 Discuta os méritos em usar entrevista pessoal, por telefone, correio ou internet como método de coleta de dados para cada uma das situações abaixo:

a. Diretor de marketing de uma rede de televisão quer estimar a proporção de pessoas no país assistindo a determinado programa.

b. Um editor quer conhecer a opinião dos leitores a respeito dos tipos de notícias do seu jornal.

c. Um departamento de saúde quer estimar o número de cachorros vacinados contra a raiva no ano passado.