

ALBERTO CARLOS ALMEIDA

COMO SÃO FEITAS AS PESQUISAS ELEITORAIS E DE OPINIÃO



17 cópias
14. cópias
ITEM 9
OBRIGATORIO
cap. 2, pp. 45-55
LEITURA

- Ostrom Jr., Charles & Simon, Dennis. Promise and performance: a dynamic model of presidential popularity. *The American Political Science Review*, 79:334-58, 1985.

São muitos os textos que analisam a suposta influência das pesquisas sobre o voto. Destaco os seguintes:

- Ansolabehere, Stephen & Iyengar, Shanto. Of horseshoes and horse races: experimental studies of the impact of poll results on electoral behavior. *Political Communication*, 11(4), 1994.
- Gelman, A. & King, Gary. Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, 1993.
- Jackson, John. Election night reporting and voter turnout. *American Journal of Political Science*, 27(4), 1983.
- McAllister, Ian & Studlar, D. T. Bandwagon, underdog, or projection? Opinion polls and electoral choice in Britain, 1979-1987. *The Journal of Politics*, 53(3), 1991.

C a p í t u l o 2

A amostra

"Eu nunca fui entrevistado em uma pesquisa de opinião"

Nas décadas de 1980 e 90, um dos políticos mais importantes do país no período, Leonel Brizola, criticava publicamente, e com frequência, as pesquisas de opinião política, perguntando aos eleitores se eles já haviam sido entrevistados em alguma pesquisa desse tipo. Como as pesquisas se baseiam em amostras, apenas um percentual muito pequeno dos eleitores é entrevistado. Por isso, a resposta mais comum à pergunta de Brizola era "não". Vejamos um cálculo simples. Se, no município de São Paulo, for realizado um total de 100 pesquisas, cada uma com mil entrevistas, e se nenhum eleitor responder a duas ou mais pesquisas, então serão entrevistados 100 mil eleitores. O eleitorado do município de São Paulo em 2000 era de aproximadamente 7 milhões de eleitores. Assim, a probabilidade de alguém ser entrevistado em uma das 100 pesquisas era de apenas 1,4%! Brizola, portanto, demonstrou que conhecia os princípios da amostragem, embora usasse esse conhecimento para criticar as pesquisas.

A pesquisa por amostragem tem por objetivo, entrevistando-se uma parcela muito pequena da população que se deseja pesquisar, realizar afirmações válidas para a população como um todo. Não é preciso entrevistar uma grande percentagem de pessoas para saber com precisão o que pensa a população pesquisada sobre determinado assunto. Já é possível, pois, adiantar uma das conclusões deste capítulo: a crítica de Brizola (e de outros políticos com críticas idênticas), além de injusta, carece de fundamento científico.

Tamanho não é documento

O caso mais famoso de amostra incorreta aconteceu em 1936 nos Estados Unidos. Naquele ano, quando foi eleito presidente Franklin D. Roosevelt, a *Literary Digest* realizou uma pesquisa de opinião, com mais de 2 milhões de entrevistados, que apontou como vitorioso o republicano Alf Landon. Roosevelt venceu em 46 dos 48 estados. Na mesma eleição, o George Gallup previu corretamente o resultado da eleição com uma amostra bem menor. Como isso pode ter ocorrido?

A resposta é simples. De nada adianta realizar milhões de entrevistas se os entrevistados não forem representativos da população. É melhor, ao contrário, realizar poucas entrevistas representativas da população. Em suma, pela menos no que se refere às amostras, tamanho não é documento.

O que é uma amostra e por que funciona

Definição de amostra

Uma parte da população, aquela parte que se selecionou para extrair a informação que se deseja obter. A amostra deve ser uma réplica em pequena escala de toda a população.

Definição de população

O grupo total de pessoas, animais, células, ou coisas sobre as quais se quer obter informações.

As duas definições possibilitam que se imaginem vários exemplos de amostras e populações. Deve-se ter em mente que, sempre que se imagina e se elabora cientificamente uma amostra, é fundamental saber o que se vai pesquisar, que informação se deseja obter.

Quadro 2
Finalidade do levantamento de dados, população e amostra

O que se quer estudar	População	Amostra
Comportamento eleitoral no município de São Paulo	Eleitores do município de São Paulo	Eleitores do município de São Paulo
Preconceito racial no Brasil	População adulta brasileira	População adulta brasileira
Avaliação que os consumidores fazem da cerveja A	Consumidores da cerveja A	Consumidores de cerveja A
O que pensam os usuários de ônibus sobre a qualidade desse serviço de transporte	Usuários de ônibus	Usuários de ônibus
O que pensam os empresários industriais argentinos sobre o Mercosul	Empresários industriais argentinos	Empresários industriais argentinos
O que os médicos do Hospital X acham que deveria ser feito para aumentar as internações neste hospital	Médicos do Hospital X	Médicos do Hospital X

O Quadro 2 é bastante ilustrativo. Note-se que a terceira coluna (amostra) é uma réplica da segunda coluna (população) — uma réplica em tamanho reduzido. Para populações superiores a 10 mil, uma amostra de 1.100 casos já significa um estudo de grande precisão.³ Assim, o que a coluna “amostra” não revela é que as amostras são sempre bem menores do que a população e, tão importante quanto isso, que existem métodos para se “ti-

³ Existem inúmeros exemplos de estudos baseados em amostras. Os exames de sangue que muitas vezes somos obrigados a fazer são estudos baseados em amostras. Quando a cozinheira avalia a qualidade do molho ou do caldo fazendo uma “provinha”, trata-se de um “estudo” baseado em uma amostra. (Ainda bem que em nenhum desses dois casos precisa-se de toda a população para se fazer um bom estudo.) Para formular a Lei da Gravidade, Newton observou uma amostra de corpos caindo, e não toda a população de corpos caindo.

rar” amostras que sejam representativas da população. Antes de passar à descrição de alguns desses métodos, vale antes procurar entender os fundamentos intuitivos da amostra, isto é, por que ela funciona, e tomar conhecimento de um pouco de vocabulário básico.

Uma das explicações mais conhecidas para o funcionamento da amostra é a frase do famoso detetive ficcional de Arthur Conan Doyle, Sherlock Holmes, em *O signo das quatro* (1999:135):

Embora o homem individual seja um enigma insolúvel, o agregado humano representa uma certeza matemática. Nunca se pode prever, por exemplo, o que fará um homem, mas é possível prever as atitudes de certo número deles. Os indivíduos variam, mas as percentagens permanecem constantes.

Um exemplo prático da constatação de Sherlock Holmes é a regularidade que separa os dias úteis dos finais de semana. Sabe-se que nos dias úteis a economia do país funcionará, as empresas, os escritórios, os órgãos públicos etc., todos abrirão e funcionarão. Sabe-se, portanto, que, na média, a maioria das pessoas comparecerá ao trabalho nos dias úteis (e inclusive que essa média é menor na segunda-feira do que nos demais dias úteis). Mas não é possível saber (nem prever) se, por exemplo, Fulano da Silva (empregado numa determinada firma) irá trabalhar. Ele pode ficar doente, ser necessário a alguém de sua família, ser vítima de algum tipo de imprevisto ou emergência. O mesmo raciocínio se aplica a qualquer indivíduo. Em resumo: nunca se pode prever o que um homem fará, mas é possível dizer com precisão o que, em média, um número deles fará.

Tarólogos, cartomantes e amostragem

Tarólogos, videntes, cartomantes, pois-de-santo etc. conhecem a lei que rege a amostragem. Perguntem a várias pessoas o que elas gostam de fazer. Elas responderão: ir à praia, ao cinema, viajar, praticar esportes e coisas assim, que se pode prever ou “adivinhar”. Perguntem também a várias pessoas que tipos de problemas tiveram em suas vidas. Na média, elas irão dizer: problemas financeiros, frustrações amorosas, mortes de entes queridos e outras coisas que conhecemos e que se aplicam também a cada um de nós. Ainda assim, cada pessoa é individualmente diferente. Dá para perceber que não é tão difícil assim ser vidente ou cartomante!

Algum vocabulário básico e os conceitos de viés e precisão

Apresento a seguir o vocabulário básico que será utilizado de agora em diante. Vale notar que não se trata de um simples glossário, mas de conceitos importantes. Passemos a eles:

- **Unidade:** qualquer indivíduo que faça parte da população.
- **Variável:** a característica das unidades sobre as quais queremos obter informações.
- **Parâmetro:** número que descreve uma característica da população; é um número que existe, mas cujo valor não se sabe na prática.
- **Estatística:** número que descreve uma amostra. O valor da estatística é obtido quando se tem uma amostra, mas muda de amostra para amostra. Geralmente, a estatística é usada para estimar um parâmetro desconhecido.⁴

É preciso prestar muita atenção em duas definições e em uma palavra: parâmetro, estatística e estimar. Elas serão muito importantes mais adiante. Igualmente importantes são as definições de viés e precisão apresentadas a seguir.

- **Viés:** desvio consistente, repetido, e na mesma direção, da estatística amostral em relação ao parâmetro da população.
- **Precisão ou eficiência:** quando os valores de amostras repetidas da mesma população ficam muito próximos do parâmetro da população.

Um exemplo exagerado de viés é a situação na qual um pesquisador deseja mensurar a intenção de voto de uma cidade qualquer, mas só entrevista simpatizantes de um determinado partido, deixando de fora da amostra os simpatizantes dos demais partidos e aqueles que não simpatizam com partido algum. Esse é um exemplo exagerado, mas existem situações mais sutis nas quais ocorre viés. Uma delas diz respeito aos entrevistados de certas ocupações. Imagine uma situação na qual os camelôs sejam entrevistados em maior número do que sua proporção na população. Caso esse grupo ocupacional vote em massa em um candidato específico, provavelmente o

⁴ Estatística é também um corpo de conhecimento, uma disciplina, mas essa definição não nos interessa no momento.

resultado da pesquisa será enviesado. O antídoto para o viés é simples: a amostra deve ser aleatória. Isso assegura que nenhum grupo social, ou equivalente, ficará sobre-representado ou sub-representado na amostra. Assim, o que todo pesquisador e coordenador de pesquisa busca é um estudo sem viés.

Para compreender a precisão, vamos fazer um exercício hipotético. Imagine que sejam feitas 10 pesquisas com a mesma amostra, o mesmo questionário e equipes treinadas nos mesmos procedimentos. Imagine também que essas 10 pesquisas sejam realizadas no mesmo dia. É razoável supor que os resultados não sejam idênticos, mas um pouco diferentes, posto que em cada uma das pesquisas os entrevistados serão diferentes. Assim, quanto mais discrepantes entre si forem os resultados, menor a precisão das pesquisas, e quanto mais próximos entre si, maior a precisão. Se o candidato A tem 40% das intenções de voto na população (parâmetro, suponha que seja possível conhecê-lo por outros meios que não o resultado eleitoral), e se 10 pesquisas obtiverem os resultados 41, 42, 40, 43, 37, 39, 38, 40, 39 e 41, estas pesquisas serão mais precisas do que outras 10 que obtiverem 47, 32, 40, 35, 45, 41, 30, 50, 40 e 39. Na prática, o pesquisador só tem o resultado de uma pesquisa. O primeiro conjunto de pesquisas oferece resultados mais próximos do parâmetro da população do que o segundo conjunto.

Outro exemplo interessante é o de pesquisa pela Internet. As pesquisas feitas pela Internet têm necessariamente viés, ou porque os que possuem computador e acesso à rede têm um perfil socioeconômico e um modo de pensar diferentes da totalidade da população, não sendo portanto representativos dela, ou porque os que respondem voluntariamente às pesquisas o fazem por terem motivações diferentes daqueles que optam por não responder. O resultado da pesquisa é enviesado, mesmo que se aumente bastante o tamanho da amostra. Ou seja, amostras maiores apenas aumentam a precisão de uma estatística enviesada.

Quanto maior a amostra (sendo tudo o mais constante), maior a precisão. Há, todavia, um tamanho de amostra a partir do qual aumentar o número de entrevistas não oferece ganhos substanciais adicionais de precisão, considerando-se os custos crescentes associados ao aumento do tamanho da amostra.

A melhor (e mais utilizada) maneira de expressar as noções de viés e eficiência é por meio de alvos. Os quatro alvos a seguir mostram as quatro combinações possíveis entre viés e eficiência. O centro do alvo é o parâmetro da população.

1. Sem viés (ou viés baixo) e precisão elevada.

É o que todo pesquisador procura fazer. Um bom estudo realiza esses objetivos.



3. Viés baixo (ou sem viés) e pouca precisão.



4. Viés elevado e pouca precisão.



Note-se que a situação 1 é a única que deve ser buscada por pesquisadores e coordenadores de pesquisas. Como na prática tem-se apenas uma pesquisa, qualquer um dos resultados das pesquisas hipotéticas da situação 1 ficará bem próximo do parâmetro da população. Essa afirmação não é válida para as demais situações. Nas situações 2 e 4, qualquer resultado ficará distante do parâmetro da população; e na situação 3 — caso a sorte esteja do lado do pesquisador — uma de várias pesquisas pode ser a que se aproxime do parâmetro da população. Mas, como as pesquisas são realizadas para se evitar a sorte e para buscar fundamento científico para a tomada de decisões, ninguém consideraria essa possibilidade.

Para concluir, chamamos atenção para o fato de que todas as críticas realizadas às pesquisas de opinião e às pesquisas de mercado são tecnicamente

baseadas ou no conceito de viés, ou na noção de precisão, ou em ambos. Quando se diz que um instituto de pesquisa errou, está-se afirmando que houve ou viés, ou falta de precisão, ou as duas coisas simultaneamente.

Os diferentes tipos de amostra

Existem duas grandes famílias de amostras: as probabilísticas e as não-probabilísticas. Amostras probabilísticas são aquelas nas quais todas as unidades de uma população têm alguma chance de ser incluídas na amostra. Nas amostras não-probabilísticas isso não acontece: há unidades da população que não têm chance de fazer parte da amostra. Nesse tipo de amostra, os entrevistados são selecionados de acordo com algum critério predefinido.

Amostra probabilística simples

A amostra probabilística simples é selecionada por sorteio, de tal forma que cada unidade da população tenha uma chance conhecida e igual de ser sorteada. Um requisito para a seleção dessa amostra é a listagem de todas as unidades da população. De posse dessa lista, é realizado o sorteio. A imagem popular desse tipo de amostra é a de distribuir um número diferente a cada pessoa de uma população, colocar todos os números num chapéu e sortear-los. Utilizando-se a notação convencional, onde n representa o número de unidades da amostra e N o número de unidades da população, a probabilidade de seleção de cada unidade é dada por:

$$p = n/N \quad (1)$$

Assim, se em uma cidade o eleitorado for de 200 mil eleitores e a amostra for de 400 unidades, então a probabilidade de seleção é de 400/200 mil, o que é igual a 0,2%. O exemplo é apenas didático, posto que, na maioria das vezes, essa é uma forma de seleção inviável. Primeiro, porque nunca haverá uma lista completa e atualizada dos eleitores. Segundo, porque, se existisse tal lista, depois de sortear os que seriam entrevistados seria economicamente inviável realizar entrevistas pessoais e mesmo telefônicas (não só por causa da cobertura telefônica, mas também da dificuldade de se obter os números de telefone de todos os sorteados). Esse tipo de seleção é muitas

vezes viável quando a pesquisa é realizada junto a empregados de uma firma ou funcionários de um órgão público, e se tem acesso à lista de empregados ou de funcionários que formam a população a ser pesquisada. Mesmo nesse caso, depois de realizado o sorteio, a pesquisa é em geral economicamente viável quando os sorteados podem ser entrevistados por telefone e tal procedimento não prejudica a cientificidade da pesquisa.

A amostra probabilística simples é, em geral, representativa da população que se quer pesquisar. Por representativa deve-se entender a amostra que expressa as características socioeconômicas, demográficas, de opinião etc. da população. A rigor, pode acontecer que uma amostra probabilística simples não seja representativa da população. Por exemplo, em uma população formada por 50% de mulheres e 50% de homens, o sorteio pode resultar numa amostra de 65% de mulheres e apenas 35% de homens. Apesar de improvável, isso pode acontecer. A amostra continua sendo aleatória, mas não é representativa.

Antes de passarmos a outra modalidade de amostra probabilística, convém fazer algumas observações importantes sobre a amostra probabilística simples. Primeiro, em uma amostra probabilística simples, uma unidade, depois de sorteada, pode ser elegível ou não para os sorteios seguintes. Quando é elegível, trata-se de uma *amostra com reposição*, e quando não é, trata-se de uma *amostra sem reposição*. A amostra com reposição tem características estatísticas superiores à amostra sem reposição, mas, do ponto de vista prático, uma unidade pode ser selecionada mais de uma vez. No caso de populações muito grandes, as amostras podem ser tratadas como amostras com reposição, nas quais as chances de sortear duas vezes a mesma unidade são muito pequenas.

Segundo, a amostra probabilística simples é a que serve de base para se estimar as margens de erro e os intervalos de confiança da maioria das pesquisas de opinião realizadas e divulgadas pela mídia. Terceiro (e aplicável a todas as amostras probabilísticas), a seleção das unidades da amostra em nada depende de escolhas feitas pelos pesquisadores, mas apenas de um sorteio.

Amostra sistemática

A amostra sistemática é uma variante da amostra probabilística. Obriga que o sorteio das unidades a serem selecionadas para a amostra seja feito considerando-se todas as unidades, do início ao fim da lista.

O procedimento é simples. O primeiro passo é embaralhar a lista de unidades da população. Em seguida, define-se o intervalo de seleção das unidades da amostra dividindo-se N por n . Por exemplo, se a amostra é de mil unidades e a população de 100 mil, o salto será de 100 mil/mil, ou seja, a cada 100. Em seguida sorteia-se o ponto de partida entre 1 e 100. Suponhamos que o número sorteado tenha sido 47. A partir de então seleciona-se a centésima unidade (147, 247, 347 etc.) até percorrer toda a lista. O último número sorteado seria 99.947. O sorteio de um ponto de partida para a contagem do intervalo é de suma importância porque assegura que todas as unidades têm chances de cair na amostra.

A amostra sistemática, além de ser aleatória, tem a vantagem de sempre ser representativa quando o intervalo de seleção não está correlacionado com alguma característica da listagem de unidades da população. Ou seja, considerando-se o exemplo acima, se a cada 100 unidades houver uma pessoa que tenha uma característica X , então a amostra não será representativa da população caso esta tenha (em igual proporção), além da característica X , também as características Y e Z .

Amostra estratificada

A amostra estratificada é outra modalidade de amostra probabilística que procura assegurar que a amostra seja realmente representativa da população. É realizada quando se divide a população em subconjuntos ou estratos e, dentro de cada estrato, se sorteia uma amostra aleatória, simples ou sistemática.

Imagine-se, por exemplo, uma pesquisa que tenha como população todos os deputados estaduais das assembleias legislativas dos estados brasileiros. Em vez de se sortear uma amostra aleatória simples ou uma amostra sistemática, a população de deputados estaduais poderia ser dividida nos seguintes estratos: partido político (partido de direita, centro e de esquerda) e região do país (as cinco grandes regiões). Essa estratificação cria 15 subcategorias: partido de direita e cada grande região (cinco categorias), partido de centro e cada grande região (cinco categorias) e partido de esquerda e cada grande região (cinco categorias). Feito isso, seria sorteada uma amostra em cada uma dessas subcategorias.

A principal vantagem da amostra estratificada é assegurar, caso seja do interesse do pesquisador ou de quem contrata a pesquisa, a representatividade de cada subcategoria. Desse modo é possível comparar as opiniões

dos deputados estaduais de partidos de esquerda do Nordeste com as dos deputados estaduais de partidos de esquerda do Sudeste ou do Sul.

Note-se que a divisão em estratos exige que o coordenador da pesquisa tenha informação teórica sobre que variáveis são relevantes para o estudo. Neste exemplo considera-se que o posicionamento do partido político no espaço ideológico direita-esquerda e a região do país são variáveis relevantes para explicar o comportamento ou a opinião dos deputados estaduais. Todavia, uma dessas duas variáveis pode não ser de fato relevante, ou ainda estar faltando algum outro estrato, como o quão veterano é o deputado no exercício do poder legislativo (mensurado em número de mandatos). Assim, uma possível desvantagem da amostra estratificada é que o pesquisador necessita, para tirar a amostra, de informações prévias sobre todos os indivíduos da população. Se tais informações inexistirem será impossível elaborar a amostra.

Amostra por conglomerado e em múltiplos estágios

A amostra por conglomerado permite fazer várias entrevistas numa mesma unidade geográfica. Em geral é parte de um desenho de amostra em múltiplos estágios.

Imagine-se, por exemplo, que se deseje realizar uma pesquisa junto à população adulta brasileira. Uma forma de fazer essa amostra é sortear, em múltiplos estágios, áreas geográficas e, não, indivíduos (note-se que nas modalidades de amostra apresentadas antes sortearam-se indivíduos e, não, áreas geográficas). Primeiro, listam-se todos os municípios do Brasil e realiza-se um sorteio de, por exemplo, 100 municípios. Como as populações são bastante diferentes e os municípios mais populosos concentram uma enorme proporção da população brasileira, esse sorteio pode ser feito com *probabilidade proporcional ao tamanho* (PPT), isto é, quanto maior o município, maiores as chances de ele ser selecionado para a amostra. O município, portanto, será o que se denomina Unidade Primária de Amostragem (UPA). Em seguida, devem ser listados todos os setores censitários⁵ dos 100 municípios

⁵ O IBGE divide o país em milhares de setores censitários. Cada setor censitário é uma área geográfica contígua, com um número determinado de domicílios. Em áreas urbanas, por exemplo, é comum um setor censitário coincidir com um quarteirão ou bloco. Há casos em que um prédio constitui um setor censitário (quando, por exemplo, nele há um elevado número de domicílios). Nas áreas rurais os setores são em geral muito extensos, de modo a conter um número razoável de domicílios.

selecionados. Para cada município é sorteado um determinado número de setores censitários. Em seguida, para os setores censitários sorteados, listam-se os domicílios neles existentes e que tenham população residente, e sortie-se um certo número de domicílios nos quais os entrevistadores terão que realizar entrevistas.

Cabe ressaltar que, até esse estágio de sorteio, não se falou em pessoas ou indivíduos, mas em unidades territoriais ou físicas: municípios, setores censitários e domicílios. Numa amostra por conglomerado, o entrevistador, chegando ao domicílio, tem que seguir uma regra para selecionar a pessoa a ser entrevistada. É importante salientar que não compete ao entrevistador decidir que pessoa deve ser entrevistada; ele deve apenas seguir a regra de seleção já definida pela coordenação da pesquisa. Uma maneira de fazer essa seleção é arrolar os moradores de cada domicílio sorteado pela data de nascimento e entrevistar aquele cuja data estiver mais próxima da data de início da pesquisa. Assegura-se assim a aleatoriedade no sorteio do entrevistado.

Note-se que as características dos indivíduos a serem entrevistados — e também dos que compõem a população pesquisada (no nosso exemplo, a população adulta brasileira) — não precisam ser conhecidas nem antes nem durante o processo de seleção da amostra. As únicas informações necessárias referem-se às unidades geográficas, e são mais fáceis de obter.

Amostra por cotas

Todas as modalidades de amostra já apresentadas pertencem à grande família das amostras probabilísticas. A amostra por cotas, ao contrário, pertence ao grupo das amostras não-probabilísticas.⁶ Na realidade ela é o principal tipo de amostra não-probabilística, sendo muito utilizada no Brasil na realização de inúmeras modalidades de pesquisa, entre as quais se destacam as pesquisas de opinião política e de intenção de voto.

⁶ Existem outras amostras não-probabilísticas além das tratadas aqui. São elas: amostra conveniente, casos mais semelhantes/mais diferentes, casos típicos, casos críticos e bola de neve. Esta última modalidade leva esse nome porque cada unidade selecionada da amostra indica a seguinte.

Na amostra por cotas é necessário dividir a população em subgrupos — como homem e mulher, branco e negro, escolaridade alta e escolaridade baixa, jovens, adultos e idosos — e calcular o tamanho proporcional de cada subgrupo. Em seguida, é preciso definir o número total de entrevistas a serem feitas e dividi-las de acordo com as proporções encontradas para cada um dos subgrupos da população. Assim, por exemplo, se na população a ser estudada há 53% de mulheres e 47% de homens, e se o número total de entrevistas é de 400, então deverão ser entrevistadas 212 mulheres (53% da amostra) e 188 homens (47% da amostra). A mesma lógica se aplica aos demais subgrupos e também a cruzamentos de subgrupos. Ou seja, se existem 33% de mulheres de escolaridade baixa na população, este deve ser o percentual de mulheres de escolaridade baixa na amostra, e assim por diante nas demais combinações de subgrupos.

Na seleção da pessoa a ser entrevistada, o entrevistador deve escolher aquela que preencher as características da cota predeterminada e que ele terá que cumprir quando estiver coletando os dados. Por exemplo, um entrevistador pode ter que entrevistar cinco mulheres de escolaridade baixa e quatro de escolaridade alta. Não importa que pessoa será escolhida para dar a entrevista. Basta que tenha as características definidas na cota. No final da coleta de dados, ao se somar o trabalho de todos os entrevistadores, a amostra terá as mesmas proporções da população no que tange às variáveis escolhidas para definir as cotas.

Uma das vantagens da pesquisa por cotas é o baixo custo aliado à rapidez. Há, porém, duas desvantagens importantes. Uma delas é que o entrevistador seleciona o entrevistado, o que pode resultar em viés. Para que isso ocorra, apesar de cumpridas corretamente as cotas da amostra, basta que se verifique o seguinte: a) haver outra característica da população que não faça parte da cota, mas que esteja correlacionada com a informação que se deseja obter; e b) que os entrevistadores sistematicamente entrevistem mais pessoas com essa característica. Por exemplo, se a cor não fizer parte da cota, mas apenas sexo, idade e escolaridade, e se os brancos votarem de forma bem diferente dos negros, muito provavelmente o resultado da pesquisa apresentará viés se todos os entrevistados forem negros (ou brancos).

A outra desvantagem é que, como o entrevistador sempre cumpre a cota, tende a haver uma sub-representação na amostra das pessoas difíceis de ser encontradas. Note-se que essa falha pode ocorrer até mesmo numa

amostra probabilística, desde que um determinado número de unidades sorteadas não seja facilmente encontrado e não se insista em entrevistar essas pessoas.

Tamanho da amostra, margem de erro e intervalo de confiança

Uma das informações mais disseminadas sobre as pesquisas de opinião é a *margem de erro*. Com menor visibilidade do que a margem de erro, provavelmente porque seu entendimento é menos intuitivo, o *intervalo de confiança* é também uma informação muito divulgada. Todavia, em ambos os casos, os consumidores de pesquisas e o grande público de modo geral compreendem apenas parcialmente as duas noções.

"Margem de erro" e "intervalo de confiança" são duas noções conectadas e, por isso, precisam ser explicadas e compreendidas em conjunto. Vejamos de maneira breve as duas definições:

- ☐ *Margem de erro*: diz o quão perto a estatística da amostra cai ou está em relação ao parâmetro da população.
- ☐ *Intervalo de confiança*: diz que percentual de todas as amostras possíveis satisfaz a margem de erro.

Assim, quando se afirma que a margem de erro é de três pontos percentuais para cima e para baixo e que o intervalo de confiança é de 95%, está-se afirmando que, se na amostra um candidato tiver 30% das intenções de voto (lembre-se, esse número é a estatística amostral), na população esse candidato deve ter entre 27 e 33% das intenções de voto. Além disso, como o intervalo de confiança é de 95%, uma em cada 20 pesquisas feitas com a mesma metodologia possivelmente irá apresentar um resultado fora da margem de erro. Em outras palavras, o erro de uma pesquisa, entre 20 realizadas (isso é uma probabilidade), será maior do que três pontos percentuais para cima ou para baixo.

Cumprir a atenção para alguns elementos relevantes desse exemplo. Primeiro, o erro amostral e o intervalo de confiança são os instrumentos que permitem fazer uma estimativa. O exemplo acima é uma estimativa do parâmetro da população baseada na estatística amostral. Não se deve esquecer que o parâmetro é sempre um valor desconhecido. Só as pesquisas eleitorais podem ser conferidas (ter sua medição validada, no

jargão técnico) mediante comparação com os resultados da eleição; e, mesmo assim, apenas as pesquisas de boca-de-urna ou aquelas feitas imediatamente antes do pleito. É a chamada validação externa.

Segundo, a existência de intervalos de confiança (que podem ser maiores ou menores, mas sempre existirão) é a admissão de que a ciência pode falhar e de que a probabilidade de ocorrência dessa falha pode ser estimada. A rigor, o intervalo de confiança é a maneira científica de fazer a seguinte afirmação: mesmo realizando-se pesquisas totalmente corretas, coordenadas por pessoas absolutamente honestas e qualificadas do ponto de vista técnico, ainda assim a pesquisa pode apresentar um resultado significativamente errado. A probabilidade de isso acontecer é pequena, mas existe e efetivamente ocorre.

Tercero, convém sublinhar que estamos tratando apenas do erro amostral. É comum que cidadãos e figuras públicas insatisfeitas com os resultados de pesquisas proponham o aumento do tamanho da amostra visando a reduzir a margem de erro. Tais propostas são incompletas e insuficientes, posto que reduziram apenas o erro amostral, justamente a margem de erro que acabamos de definir. Ocorre que o erro não-amostral (que será tratado nos capítulos 3 e 4) pode levar a resultados muito mais distantes do parâmetro da população do que o erro amostral. É também que é um erro muito mais difícil de ser detectado e controlado. Em suma, o erro amostral é apenas parte do erro da pesquisa — muitas vezes a parte menos grave e menos problemática.

Outras maneiras de se dar a mesma informação

No exemplo dado anteriormente, no qual a pesquisa tem uma margem de erro de três pontos percentuais, intervalo de confiança de 95% e o candidato tem 30% das intenções de voto, essa mesma informação pode ser dada de várias maneiras:

- ☐ Em 95% de todas as amostras possíveis, a estatística amostral assumirá valores em um intervalo de mais ou menos três pontos percentuais do parâmetro da população.
- ☐ A pesquisa tem uma margem de erro de mais ou menos três pontos percentuais e 95% de confiança.
- ☐ Com 95% de confiança, a proporção de todos os eleitores que irão votar no candidato encontra-se no intervalo estatístico de 0,30 +/- 0,03.

A margem de erro e o intervalo de confiança de uma pesquisa são, em grande parte, funções do tamanho da amostra. Se fixarmos o intervalo de confiança em 95% e aumentarmos o tamanho da amostra, a margem de erro diminuirá (algo desejável). Mas se fixarmos a margem de erro em três pontos percentuais e aumentarmos o tamanho da amostra, o intervalo de confiança aumentará (algo também desejável). A tabela 1 mostra, para três diferentes intervalos de confiança (99, 95 e 90%), as respectivas margens de erro, de acordo com o tamanho da amostra. O gráfico 1 apresenta os dados da tabela, com a vantagem de tornar mais intuitivos os ganhos marginais decrescentes na redução da margem de erro em função do aumento da amostra.

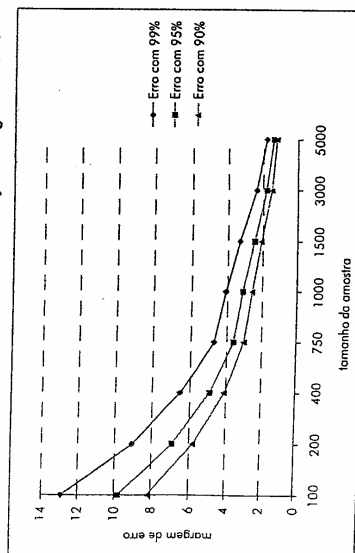
Tabela 1

Tamanho da amostra, intervalo de confiança e margem de erro

Tamanho da amostra	100	200	400	750	1.000	1.500	3.000	5.000
Erro com 99%	12,9	9,1	6,5	4,7	4,1	3,3	2,4	1,8
Erro com 95%	9,8	6,9	4,9	3,6	3,1	2,5	1,8	1,4
Erro com 90%	8,2	5,8	4,1	3,0	2,6	2,1	1,5	1,2

Gráfico 1

Tamanho da amostra, intervalo de confiança e margem de erro



Pode-se perceber que são expressivos os ganhos em termos de redução da margem de erro de uma amostra de tamanho 100 até uma amostra de tamanho 750, para quaisquer dos intervalos de confiança apresentados. A partir de 750, esses ganhos são bem menores. Isso sugere que, considerando-se o aumento de custos associado ao aumento do número de entrevistas, há um determinado tamanho de amostra a partir do qual deixa de ser proveitoso continuar aumentando a amostra.

Os números acima também deixam claro que é possível ter diferentes combinações de margem de erro e intervalos de confiança para um mesmo número de entrevistas. Tomemos uma amostra de tamanho 400. Neste caso, o erro para 99% de confiança é 6,5 pontos percentuais, 4,9 para 95% de confiança e 4,1 para 90% de confiança. Assim, um coordenador de pesquisa inescrupuloso pode adaptar aos seus objetivos, para um número fixo de entrevistas, a margem de erro que melhor lhe convier — certamente aquela que lhe permitir afirmar que sua pesquisa ficou dentro da margem de erro.

Vale notar que, nas amostras grandes, as diferenças entre as margens de erro para cada intervalo de confiança são menores do que nas amostras pequenas. Mesmo assim, a margem de erro de uma pesquisa com intervalo de confiança maior (99%, por exemplo), para o mesmo tamanho de amostra será sempre maior do que a margem de erro de uma pesquisa com um intervalo de confiança menor (95%, por exemplo).

Para concluir, convém reiterar um ponto importante. As margens de erro e os intervalos de confiança aqui apresentados são válidos para a amostra probabilística simples. Para outros tipos de amostra probabilística, é comum ser necessário corrigir as margens de erro (para os mesmos intervalos de confiança apresentados) em função do efeito do desenho da amostra (*design effect*). Em amostras não-probabilísticas, não é possível calcular margens de erro e intervalos de confiança, pode-se apenas estimar empiricamente os seus valores.

Algumas dicas para entender as afirmações sobre margem de erro e grau de confiança

- As conclusões acerca da margem de erro e do intervalo de confiança aplicam-se sempre à população, jamais à amostra.

continua

- ☐ Assim sendo, as conclusões acerca da população nunca são totalmente certas, isto é, nunca há certeza absoluta.
- ☐ A imprensa costuma divulgar apenas a margem de erro. Quando isso acontece, supõe-se um intervalo de confiança de 95%.
- ☐ Quando se deseja uma margem de erro menor com o mesmo intervalo de confiança, deve-se aumentar o tamanho da amostra.
- ☐ Quando se quer 99% de intervalo de confiança, deve-se acelar uma margem de erro maior do que a utilizada para 95% de confiança. Há uma compensação entre margem de erro e intervalo de confiança.

Amostras na prática

Vejam os agora, passo a passo, a elaboração de duas amostras (pelo DataUff) para duas pesquisas diferentes. A primeira é a amostra para o município do Rio de Janeiro que o DataUff utilizou nas pesquisas de opinião política e intenção de voto fornecidas para o *Journal da Brasil* e a America Online no ano 2000. É uma amostra por cota; portanto, não-probabilística. A segunda amostra é probabilística, de abrangência estadual, e foi elaborada para uma pesquisa acadêmica sobre relações raciais contratada pelo Centro de Estudos e Apoio às Populações Marginalizadas (Ceap) e financiada pela Fundação Ford.

Amostra por cotas

Como já se viu, a amostra por cotas é não-probabilística. Esse é o tipo de amostra utilizado por todos os grandes institutos de pesquisa para realizar pesquisas eleitorais. Assim, apesar das variações entre os desenhos amostrais de cada instituto, se a cota for utilizada em apenas uma etapa da seleção dos entrevistados, já não se pode mais falar em amostra probabilística.

Primeiro vamos arrolar todas as etapas a serem cumpridas para a elaboração da amostra. Em seguida, cada etapa será tratada de modo mais detalhado. Para obter uma amostra por cotas deve-se dar os seguintes passos:

1. Definir que variáveis são relevantes para o estudo planejado; elas formarão as cotas. No caso de pesquisas eleitorais, as variáveis relevantes são sexo, idade e escolaridade, além de local de moradia e município em que o eleitor vota.

2. Obter os dados censitários com os números absolutos referentes às variáveis escolhidas. Em geral, esses dados são produzidos pelo IBGE.
3. Calcular as proporções das variáveis para a população. No caso das pesquisas eleitorais, a população pesquisada são os eleitores de um município, de um estado, ou do país.
4. Definir o tamanho da amostra.
5. Multiplicar as proporções de cada variável obtida, ou o cruzamento destas, pelo tamanho da amostra. Assim se obtém o número de entrevistas por cota.

Passo 1: Definir as variáveis relevantes para o estudo planejado: são elas que formam as cotas.

O pesquisador deve sempre buscar uma amostra que seja representativa da população que ele quer pesquisar. Assim, como a população é o eleitorado do município do Rio de Janeiro, para que uma amostra seja representativa desse eleitorado ela tem que apresentar as mesmas proporções de sexo, faixa etária e nível de escolaridade existentes na população. Além disso, essas proporções devem ser desagregadas por unidades geográficas. No caso desta amostra, elas serão os distritos da cidade, que têm o nome de regiões administrativas, segundo a Prefeitura do Rio de Janeiro. Se, por exemplo, forem realizadas entrevistas no distrito de Copacabana (atenção, distrito é geograficamente maior do que bairro), deve-se buscar a seguinte informação: quantas mulheres de 25 a 34 anos existem no distrito de Copacabana como proporção de todo o eleitorado da cidade. O mesmo se aplica aos níveis de escolaridade, que podem ser cruzados ou não com sexo e idade. No caso da amostra aqui discutida, a cota de escolaridade é controlada em separado de sexo e idade, e estes, sim, é que são cruzados.

Note-se que como no Brasil o voto é obrigatório e universal, supõe-se que a população acima de 18 anos seja toda composta de eleitores. O voto só é facultativo para analfabetos, pessoas de 16 e 17 anos e pessoas com mais de 70 anos. No momento da entrevista, pergunta-se ao entrevistado se irá ou não votar nas eleições. Se a resposta for não, obviamente não se faz a entrevista. Nos países onde o voto não é obrigatório, é necessário estimar a probabilidade de abstenção dos diferentes segmentos do eleitorado.

Passo 2: Obter os dados censitários com os números absolutos referentes às variáveis escolhidas.

O IBGE disponibiliza ao público os dados referentes a sexo, idade e escolaridade de todos os municípios do Brasil, desagregados pelos diferentes distritos de cada município. Na contagem da população de 1996, os distritos do município do Rio de Janeiro eram 26, mas uma vez que o distrito de Paqueta tinha na época um eleitorado muito pequeno como proporção do eleitorado total da cidade, foram utilizados apenas 25 distritos, como mostra o quadro 3:

Quadro 3
Distritos do município do Rio de Janeiro

Zona Portuária	São Cristóvão	Méier	Santa Cruz
Lagoa	Inhaúma	Compo Grande	Pavuna
Penha	Bangu	Barra	Copacabana
Jacarepaguá	Santa Teresa	Botafogo	Ramos
Anchieta	Rio Comprido	Vila Isabel	Madureira
Centro	Tijuca	Itaquá	Ilha do Governador
			Guaratiba

Fonte: IBGE.

É importante registrar que muitos distritos têm os mesmos nomes de vários bairros, mas são mais abrangentes que os bairros que levam o mesmo nome. Por outro lado, muitos bairros que não aparecem na lista de distritos estão incluídos em algum distrito da cidade. Por exemplo, o bairro de Laranjeiras está contido no distrito de Botafogo. Resumindo: toda a cidade do Rio de Janeiro está contida nestes 25 distritos, excetuando-se Paqueta.

Para cada distrito da cidade há dados absolutos de sexo e idade, e de escolaridade. O maior distrito do município é Bangu, e tem o seguinte perfil etário por sexo:

Tabela 2
Número de indivíduos do distrito de Bangu por sexo e faixas de idade

Sexo/Idade	16 a 17	18 a 24	25 a 34	35 a 44	45 a 59	60 ou mais
Homem	10.791	37.552	51.466	40.105	35.908	20.808
Mulher	11.022	37.490	55.060	45.073	41.301	29.294

Fonte: IBGE.

No total, Bangu tem 415.870 eleitores, o que corresponde a aproximadamente 10% do eleitorado do município do Rio de Janeiro.⁷ Dos eleitores residentes em Bangu que estão na faixa dos 25 a 34 anos de idade, 55.060, por exemplo, são mulheres e 51.466 são homens. Além desse dado, é necessário obter, para o mesmo distrito, o número absoluto de pessoas pertencentes a cada nível de escolaridade.

Tabela 3

Número de indivíduos do distrito de Bangu por nível de escolaridade

Sem instrução	Primário completo	1º grau completo	2º grau completo	Superior completo
37.428	205.024	79.015	75.273	19.130

Fonte: IBGE.

De todos os eleitores de Bangu, 37.428 não têm instrução, 205.024 têm o primário completo e assim por diante. A elaboração da amostra terá que respeitar as proporções que serão obtidas dos números absolutos acima: proporção de eleitores por distrito, proporção por faixas de sexo e idade, e níveis de escolaridade. É o passo seguinte na elaboração da amostra.

Passo 3: Calcular as proporções das variáveis para a população.

Este passo do cálculo da amostra é bastante simples e direto. Obtidos todos os números absolutos, por distrito, faixas de idade, escolaridade e sexo, dividem-se esses números pelo eleitorado total do município do Rio de Janeiro que, de acordo com os mesmos dados da contagem de 1996, era de 4.003.975. Ou seja, se dividirmos o total de eleitores de Bangu pelo total de eleitores do município, encontraremos 10,38%. Da mesma maneira, os outros valores absolutos encontrados para Bangu (sexo e idade, e escolaridade), e para todos os demais distritos, são divididos pelo tamanho total do

⁷ O segundo maior eleitorado do Rio de Janeiro é do distrito do Méier, com 7,8% do eleitorado total.

eleitorado. Com isso são encontradas as proporções para todas as variáveis. As tabelas a seguir apresentam as proporções de sexo e idade, e escolaridade para Bangu, e as proporções de entrevistados que devem ser realizadas em cada distrito da cidade. Vale lembrar que será preciso fazer para cada distrito o mesmo que se fez para Bangu no que se refere às três variáveis sociodemográficas da cota.

Tabela 4

Proporção de eleitores do distrito de Bangu, por sexo e faixa etária, como proporção do eleitorado total de Bangu

Sexo/Idade	16 a 17	18 a 24	25 a 34	35 a 44	45 a 59	60 ou mais
Homem	0,026	0,09	0,124	0,096	0,086	0,05
Mulher	0,027	0,09	0,132	0,108	0,099	0,07

Tabela 5

Proporção de eleitores do distrito de Bangu, por nível de escolaridade, como proporção do eleitorado total de Bangu

Sem instrução	Primário completo	1ª grau completo	2ª grau completo	Superior completo
0,09	0,49	0,19	0,18	0,05

Tabela 6

Proporção de eleitores dos 25 distritos do município Rio de Janeiro como proporção do eleitorado total do município

Z. Pontuária	0,8	São Cristóvão	1,5	Méier	7,9	Santa Cruz	4,1
Lagoa	4,3	Inhaúma	3,7	Campo Grande	6,5	Pavuna	3,1
Penha	5,7	Bangu	10,4	Barra	1,8	Copacabana	3,6
Jacarepaguá	7,6	Santa Teresa	0,8	Bolado	5,1	Ramos	5,0
Archieta	2,5	Rio Comprido	1,5	Vila Isabel	3,8	Madureira	6,9
Centro	1,0	Tijuca	3,8	Itaí	3,9	Ilha	3,6
						Guaratiba	1,0

Passo 4: Definir o tamanho da amostra.

Para se definir o tamanho da amostra é necessário que o coordenador da pesquisa estipule que margem de erro e que intervalo de confiança desejada. Não há cálculo estatístico que permita obter o tamanho da amostra no caso de amostras não-probabilísticas. Por isso, calcula-se o tamanho para uma amostra probabilística simples e define-se que este será o tamanho da amostra por cotas. Na tabela 1, estão definidos os tamanhos de amostra para cada combinação de margem de erro e intervalo de confiança. Assim, nosso exemplo será para uma amostra de 3,1 pontos percentuais de margem de erro e um intervalo de confiança de 95%, o que significa uma amostra de mil casos.

Passo 5: Multiplicar as proporções de cada variável obtida, ou o cruzamento destas, pelo tamanho da amostra. Assim se obtém o número de entrevistados por cota.

Multiplicando-se o número de entrevistas da amostra — mil — pelas proporções obtidas para Bangu, temos os resultados apresentados na tabela 7:

Tabela 7

Número de entrevistas no distrito de Bangu por sexo e faixa etária

Sexo/Idade	16 a 17	18 a 24	25 a 34	35 a 44	45 a 59	60 ou mais
Homem	3	9	13	10	9	5
Mulher	3	10	14	11	10	7

Tabela 8

Número de entrevistas no distrito de Bangu por níveis de escolaridade

Sem instrução	Primário completo	1ª grau completo	2ª grau completo	Superior completo
9	51	20	19	5

Tabela 9
Número de entrevistas em cada um dos 25 distritos do município do Rio de Janeiro

	8	São Cristóvão	15	Méier	79	Santa Cruz	41
Z. Portuária	43	Inhaúma	37	Campos Grande	65	Pavuna	31
Lagoa	57	Bangu	104	Barra	18	Copacabana	36
Penha	76	Santa Teresa	9	Botafogo	51	Ramos	50
Jacarepaguá	25	Rio Comprido	15	Vila Isabel	38	Madureira	69
Anchieta	10	Tijuca	38	Iolô	39	Ilho	36
Centro						Guaratiba	10

O mesmo precisa ser feito para todos os distritos do município do Rio de Janeiro. Ao fim e ao cabo, quando forem somadas todas as entrevistas por faixas de sexo e idade, e por escolaridade, as proporções totais coincidirão com estas mesmas proporções referentes ao eleitorado de todo o município.

Ainda como parte da amostra, é necessário distribuir as entrevistas por pontos de coleta espalhados por toda a cidade. Esses pontos são sorteados, e cada equipe tem que realizar no mínimo 10 e no máximo 15 entrevistas por ponto. Cada distrito tem um determinado número de pontos de coleta (número este proporcional a algum número inteiro entre 10 e 15), nos quais as entrevistas só poderão ser realizadas com moradores dos bairros daquele distrito.

As amostras por cotas apresentam variações, mas variações sobre o mesmo tema. A grande diferença ocorre entre a amostra por cotas e a amostra probabilística, na qual o entrevistado é sorteado, não tendo que se enquadrar em qualquer perfil de sexo, idade, escolaridade, ou outra variável de interesse. A seguir veremos na prática esse tipo de amostra.

Amostra probabilística sem substituição

A amostra probabilística nunca é utilizada em pesquisas eleitorais. Mas pode ser muito útil em pesquisas de mercado e em pesquisas acadêmicas. Acima de tudo, é a teoria da amostra probabilística que serve de parâmetro para avaliar os demais tipos de amostra.

Primeiro, veremos as etapas a serem cumpridas para a elaboração da amostra e, em seguida, cada etapa será tratada de forma mais detalhada. Para se obter uma amostra probabilística, devem ser dados os seguintes passos:

1. Definir os objetivos da pesquisa: o que se deseja saber. Neste exemplo, a amostra é desenhada para uma pesquisa que tem por tema as relações raciais, e a população que se deseja representar é a de adultos residentes no estado do Rio de Janeiro. Também é preciso definir as variáveis de interesse para a análise. Assim pode-se saber se algum segmento da população deve ser sobre-representado no cálculo da amostra.
2. Dividir os estratos para o sorteio da amostra.
3. Definir o tamanho da amostra.
4. Definir os procedimentos de sorteio e realizá-lo.

Passo 1: Definir os objetivos da pesquisa — o que se deseja saber — e as variáveis de interesse para a análise.

O primeiro passo é óbvio, mas fundamental para a elaboração da amostra. O exemplo escolhido é o da pesquisa realizada pelo DataUff para o Centro de Apoio e Estudos das Populações Marginalizadas (Ceap), e financiada pela Fundação Ford, que teve como objetivo produzir informações sobre as opiniões e os valores da população adulta do estado do Rio de Janeiro no que tange às relações raciais — às relações existentes entre as populações de cores diferentes.

Uma vez definido o objetivo da pesquisa, pode-se afirmar que a amostra terá que representar a população adulta do estado do Rio de Janeiro. Sabe-se ainda que a opinião de brancos, pardos e pretos (segundo a classificação do IBGE) é fundamental para a pesquisa. Isso porque, no caso de relações raciais, pode haver diferença — e de fato há — entre o que pensam as pessoas de diferentes cores quanto a alguns aspectos das relações entre esses grupos sociais. Como na população do estado do Rio de Janeiro há uma proporção bastante razoável de pessoas das três cores da classificação do IBGE, uma seleção aleatória dos entrevistados provavelmente levaria às mesmas proporções na população, todas elas adequadas para a análise estatística.

Note-se que, se por exemplo os brancos representassem apenas 1% da população do estado, uma amostra aleatória de 1.200 casos tenderia a selecionar 12 brancos. Se o objetivo fosse fazer uma análise que comparasse a opinião de brancos, pardos e pretos, certamente isso seria impossível com apenas 12 casos. Assim, os brancos teriam que ser sobre-representados na amostra. Esse é um exemplo hipotético, que exagera a dificuldade de representar uma população específica, mas muito comum na realidade de uma pesquisa.

No caso da pesquisa sobre relações raciais foi necessário sobre-representar o interior do estado do Rio de Janeiro. Se a amostra fosse sorteadá considerando-se as proporções das populações do interior *versus* a população da região metropolitana (RM), o número de entrevistas realizadas no interior não permitiria fazer alguns tipos de análise estatística dos dados. A distribuição real da população é de 30% no interior e 70% na RM. Na amostra foram feitos 35% de entrevistas no interior e 65% na RM. Considerando-se o tema analisado — relações raciais —, avaliou-se que a opinião das pessoas do interior (das três cores) diferiria da opinião dos mesmos grupos sociais residentes na capital ou na RM. Para testar isso foi preciso então sobre-representar a população do interior do estado na amostra.⁸

Outra decisão importante foi garantir que, na capital, 17% dos setores censitários sorteados fossem favelas (17% da população do município do Rio de Janeiro residem em favelas). Essa distribuição tornou-se necessária porque os 17% equivalem a seis setores censitários de um total de 35 setores somente para a capital. Se não se assegurasse esse número de setores para as favelas, essas áreas poderiam ficar com menos de seis num sorteio, o que não seria desejável para a finalidade da pesquisa. Apesar desse procedimento, convém sublinhar que o número de entrevistas realizadas em favelas não foi suficiente para permitir inferência estatística para a população favelada. O objetivo foi apenas garantir que a população das favelas ficasse corretamente representada na amostra como um todo.

⁸ Quando a análise dos dados é realizada, a sobre-representação de populações e, consequentemente, a sub-representação são corrigidas por meio de ponderação. O que é sobre-representado é multiplicado por um fator menor do que 1, e o sub-representado, por um fator maior do que 1, de tal maneira que a amostra represente, sem viés, a população que se deseja estudar.

Passo 2: Definir os estratos para o sorteio da amostra.

Definem-se os estratos para o sorteio da amostra considerando-se duas coisas: a) as variáveis relevantes para explicar eventuais diferenças na opinião das pessoas; e b) as informações disponíveis para a elaboração da amostra. Caso uma dessas informações disponíveis fosse uma lista com o nome de cada adulto residente no estado do Rio de Janeiro, poder-se-ia simplesmente embalar a lista de nomes e sortear as pessoas que seriam entrevistadas com base em algum critério que assegurasse a aleatoriedade. Mas não existe uma lista assim. Por isso, deve-se recorrer a outros procedimentos para o sorteio.

Não há uma lista de nomes, mas existe uma lista de regiões geográficas, municípios, setores censitários, domicílios, e também informações sobre os setores censitários, como a escolaridade média do chefe de domicílio. Ademais, é razoável supor que a opinião das pessoas acerca de diversos temas — e o tema das relações raciais é apenas um deles — varia de acordo com algumas dessas variáveis mencionadas: pessoas que moram em municípios menos desenvolvidos pensam de maneira diferente daquelas que residem em municípios mais ricos; no mesmo município, pessoas que residem em bairros mais prósperos têm opiniões diferentes das que moram em bairros mais pobres; e o mesmo se aplica a outras variáveis.

Segundo essa lógica, os estratos escolhidos foram: regiões geográficas do estado do Rio de Janeiro, municípios classificados segundo um indicador de riqueza (PIB *per capita*) e setores censitários classificados também de acordo com um indicador de riqueza. As regiões geográficas do estado (ou regiões de amostragem) definidas foram:

- ☐ município do Rio de Janeiro (capital do estado);
- ☐ região metropolitana;
- ☐ região Serrana e das baixadas litorâneas;
- ☐ regiões Centro-Sul, Médio Paraíba e da bacia da ilha Grande;
- ☐ regiões Norte e Noroeste.

Passo 3: Definir o tamanho da amostra.

Define-se o tamanho da amostra antes de iniciar o trabalho de campo. O coordenador da pesquisa é o responsável por isso. Quando se define o tamanho da amostra, levam-se em conta os seguintes fatores: custo, interva-

lo de confiança e margem de erro da pesquisa, além do tipo de análise que se deseja realizar. Numa situação irreal na qual o custo não constituísse problema, quanto maior a amostra melhor. Contudo, o orçamento disponível para a realização da pesquisa é o primeiro condicionante, e o mais importante, para determinar o tamanho da amostra. Há casos em que os recursos são tão poucos que é melhor não fazer pesquisa alguma, pois uma amostra muito pequena resultaria numa margem de erro muito grande para se poder realizar inferências de boa qualidade.

Considerando-se os recursos disponíveis para a pesquisa de relações raciais, avaliou-se que seria possível arcar com os custos de uma amostra que fornecesse um intervalo de confiança de 95% e uma margem de erro de aproximadamente três pontos percentuais. Isso significaria uma amostra da ordem de grandeza de mil entrevistas, talvez um pouco mais. Além disso, um dos objetivos do estudo era comparar as opiniões dos moradores da capital, com as dos moradores dos demais municípios da região metropolitana e dos municípios do interior do estado do Rio de Janeiro. Para tanto, considerando-se que a amostra tende a se distribuir proporcionalmente à população dessas regiões, o desejável seria efetuar de mil a 1.200 entrevistas. Menos do que isso significaria fazer poucas entrevistas em cada uma dessas regiões, particularmente no interior, o que elevaria em muito a margem de erro para esses subgrupos, impossibilitando uma análise confiável dos dados.

No exemplo da pesquisa sobre relações raciais definiu-se que a amostra seria de 1.200 casos. Como se tratava de uma amostra sem substituição — se a pessoa sorteada no domicílio para conceder a entrevista não fosse encontrada não poderia ser substituída por outra pessoa —, foi necessário estimar a taxa de resposta (ou de recusa). Estimamos que a recusa a responder a pesquisa seria de 30%. Por recusa entende-se um somatório de fatores ou comportamentos: condomínios fechados que não permitem o acesso do pesquisador; pessoas que dificilmente são encontradas em casa, mesmo depois de vários contatos; pessoas que são encontradas mas que se recusam a permitir o acesso do pesquisador ao domicílio etc.

Para uma taxa de recusa de 30% seria preciso definir uma amostra de 1.700 casos. Assim, seriam sorteados 1.700 domicílios e uma pessoa em cada domicílio, mas 500 deles não seriam alcançados pela pesquisa, o que teria como resultado final uma amostra de 1.200 entrevistas. A estimativa de não-

resposta foi bastante precisa, sendo realizadas efetivamente 1.171 entrevistas, o que praticamente igualou a meta definida.

No caso das pesquisas com substituição sorteiam-se 1.200 casos, fazendo-se novo sorteio para as recusas. Essa modalidade de amostra é também muito utilizada, mas tende a resultar num viés mais elevado do que a amostra sem substituição.

Passo 4: Definir os procedimentos de sorteio e realizá-lo.

O último passo do desenho de uma amostra probabilística é definir como ela será sorteada. Cabe notar que, no exemplo da pesquisa sobre relações raciais, há três sorteios: o dos setores censitários dentro de cada região de amostragem, o dos domicílios dentro de cada setor censitário e, por fim, o da pessoa dentro de cada domicílio.

Como a amostra era de 1.700 casos e a taxa de não-resposta estimada era de 30%, estipulou-se que seriam feitas 17 entrevistas em cada setor censitário, a fim de se obter de fato, na média, 12 entrevistas por setor censitário. O número de entrevistas por setor censitário é um fator importante numa amostra. Para os objetivos deste livro, basta mencionar que a literatura especializada considera como regra básica que o número mais adequado, sem especificações excepcionais para uma pesquisa em particular, gira em torno de 10 entrevistas por setor.

Uma vez definidas quantas entrevistas realizar por setor (no caso 17), tem-se o número de setores censitários a serem sorteados (um total de 100 setores em todo o estado do Rio de Janeiro). A etapa seguinte é listar todos os setores por município e cada município dentro da região de amostragem, e hierarquizar os setores censitários de acordo com a escolaridade média dos chefes de domicílio. Assim, na prática, o que se tem é uma lista de todos os setores censitários do município do Rio de Janeiro, começando com o que apresenta a maior escolaridade média dos chefes de domicílio e terminando com o que tem a menor escolaridade média. O mesmo deve ser feito para todos os setores censitários dentro de seus municípios correspondentes. Em seguida, os municípios de cada região de amostragem são agrupados também hierarquicamente, do maior para o menor PIB *per capita*.

O resultado final desse processo de agrupamento é uma lista com todos os setores censitários do estado, hierarquicamente ordenados dentro de

cada município, também hierarquicamente ordenados de acordo com o critério de riqueza mencionado.

O passo final é sortear os setores censitários dentro de cada região geográfica de amostragem, o que se faz por meio de um pulo. Por exemplo, se em uma região de amostragem for necessário sortear 12 setores censitários e nessa região houver um total de 2.357 setores, então o pulo será de 196. A cada 196 setores, escolhe-se um. O primeiro da lista (aquele que dá início aos pulos) é escolhido por sorteio, mediante, por exemplo, uma tabela de geração de números aleatórios.

Sorteados os setores censitários, a etapa seguinte é sortear os domicílios, o que se realiza no campo. O procedimento é o seguinte: vai-se ao setor sorteado para contar e enumerar os domicílios, sorteia-se o primeiro número (também aleatoriamente) e, em seguida, define-se o tamanho do pulo para sortear os demais domicílios. Se for preciso sortear 17 domicílios e o setor tiver, no total, 340 domicílios, o pulo terá que ser de 20. Escolhe-se, então, um domicílio a cada 20, a partir do primeiro sorteado.

Após o sorteio dos domicílios é necessário sortear a pessoa a ser entrevistada. São vários os procedimentos para isso, mas em todos eles é preciso primeiro listar os adultos residentes naquele domicílio para depois sortear. Isso pode ser feito atribuindo um número a cada adulto e gerando um número aleatório. Outro procedimento, mais fácil de realizar no campo, e que foi utilizado na pesquisa sobre relações raciais, é anotar a data de nascimento de cada pessoa adulta (apenas dia e mês) e sortear aquela que tiver a data mais próxima — contando-se para frente ou para trás do ano (a direção não importa, tanto que o critério seja o mesmo para todos os sorteios) — do dia de início da pesquisa. A pessoa sorteada é a que deve ser entrevistada. Se ela não estiver em casa no dia do sorteio, é necessário fazer novas visitas até encontrá-la.

Como se pode notar no passo a passo da amostra por cotas e da amostra probabilística, essas duas modalidades de amostra diferem bastante. As amostras por cotas são mais simples, de tecnologia mais fácil e, por estarem associadas a pesquisas feitas na rua, resultam em pesquisas bem mais baratas. As amostras probabilísticas resultam em pesquisas mais caras, mas são as únicas que encontram sustentação teórica na literatura estatística. Isso não quer dizer que as amostras por cotas não sejam científicas e, sim, que

sua sustentação científica é apenas empírica. Já a amostra probabilística tem sustentação teórica e empírica.

O que ler sobre os assuntos tratados neste capítulo

Apenas duas das referências a seguir tratam especificamente de amostragem. As demais tratam de pesquisas de opinião de um modo geral, sendo a amostragem apenas um dos itens.

- Babbie, Earl. *Métodos de pesquisas de survey*. Belo Horizonte, UFMG, 1999.
- Doyle, Arthur Conan. *O signo dos quairns*. Porto Alegre, LP&M, 1999.
- Fowler Jr., Floyd J. *Survey research methods*. London, Sage, 1993.
- Henry, Gary T. *Practical sampling*. London, Sage, 1993.
- Silva, Nilza Nunes. *Amostragem probabilística*. São Paulo, Edusp, 1998.
- Sudman, S. *Reducing the costs of surveys*. Chicago, Aldine, 1967.
- Worcester, Robert M. *British public opinion*. Oxford, Blackwell, 1991.

Existem outros livros e artigos mais especializados sobre o assunto, com argumentos baseados na teoria estatística, mas não apresento essas referências porque fogem ao escopo deste livro.