CAPÍTULO 6
PROBABILIDADE E INFERÊNCIA ESTATÍSTICA

RESUMO:

Pesquisadores aspiram a tirar conclusões sobre a população inteira de casos relevantes para uma determinada pergunta de pesquisa. Contudo, na maioria das vezes, eles possuem dados para apenas uma amostra da população. Neste capítulo, apresentamos os fundamentos para fazer inferências sobre uma população enquanto observamos apenas uma amostra. Para fazer isso, nos baseamos na teoria probabilística, que introduzimos neste capítulo com extensas referências a exemplos. Concluímos o capítulo com um exemplo conhecido dos estudantes de ciência política – nominalmente, os erros a mais e a menos das pesquisas de aprovação presidencial e como eles ilustram os princípios da construção de pontes entre amostras e a população de interesse subjacente.

Como nos atrevemos a falar de leis do acaso? Não é o acaso a antítese de toda lei? – Bertrand Russel.

6.1 POPULAÇÕES E AMOSTRAS

No capítulo 5, aprendemos como mensurar nossos conceitos de interesse e como usar estatísticas descritivas para sumarizar grandes quantidades de informação sobre uma única variável. Em particular, você descobriu como caracterizar uma distribuição a partir de medidas de tendência central (como a média ou a mediana) e de medidas de dispersão (como o desvio-padrão ou o IIQ). Você pode implementar essas formu-

lações, por exemplo, para caracterizar a renda nos Estados Unidos ou as notas de uma prova que seu professor tenha acabado de divulgar.

Mas agora é o momento de estabelecer uma distinção crítica entre dois tipos de bancos de dados que cientistas sociais podem utilizar. O primeiro tipo de dado é o da **população** – isto é, os dados para todos os casos possivelmente relevantes. Um exemplo de dado populacional que pode surgir na sua cabeça em um primeiro momento é o do Censo americano, que consiste em uma tentativa do governo dos Estados Unidos de garantir a coleta de alguns dados sobre toda a população americana uma vez a cada dez anos[1]. Apesar de existirem bancos que almejam ter todos os casos da população relevante, é relativamente raro que cientistas sociais façam uso de dados pertencentes a toda a população[2].

O segundo tipo de dados consiste em uma **amostra**. Em razão da proliferação de pesquisas de opinião pública, muitos de vocês podem assumir que a palavra "amostra" implica uma "**amostra aleatória**"[3]. Mas este não é o caso. Pesquisadores podem produzir uma amostra por meio da aleatoriedade – isto é, cada membro da população tem uma probabilidade igual de ser selecionado para a amostra. Porém, as amostras podem também ser não aleatórias; quando isso ocorre, as denominamos de amostra de conveniência.

A vasta maioria das análises conduzidas por cientistas sociais são feitas com dados amostrais, não com dados populacionais. Por que fazer essa distinção? Embora a maioria esmagadora dos bancos de dados em ciências sociais seja composta de amostras, não de populações, é crítico observar que não estamos interessados nas propriedades da amostra *per se*; estamos interessados na amostra apenas na medida em que ela nos ajuda a entender uma população subjacente. Com efeito, tentamos construir pontes entre o que sabemos sobre uma amostra e o que acreditamos, probabilisticamente, ser verdade sobre a população. Esse processo é chamado de **inferência estatística**, porque utilizamos o que *sabemos* ser verdade sobre uma coisa (a amostra) para *inferir* o que é provável que seja verdade sobre outra coisa (a população).

Existem implicações sobre o uso de dados amostrais para aprender sobre populações. A primeira, e mais direta, é que esse processo de inferência estatística envolve, por definição, algum grau de incerteza. Essa noção é relativamente direta: sempre que

---

1 O *site* do Censo americano é: <http://www.census.gov> . [Para o Censo do Brasil, ver: <http://censo2010.ibge.gov.br/> (N.T.)]

2 Apesar disso, tentamos fazer inferências sobre alguma população de interesse, e está nas mãos do pesquisador definir explicitamente qual é essa população de interesse. Algumas vezes, como no caso do Censo dos Estados Unidos, a população relevante – todos os residentes nos Estados Unidos – é fácil de entender. Outras vezes, a definição da amostra é um pouco menos óbvia. Considere, por exemplo, um *survey* pré-eleição, no qual o pesquisador precisa decidir se a população de interesse são todos os adultos, ou os prováveis eleitores, ou algum outro grupo.

3 Quando discutimos o desenho de pesquisa no capítulo 4, distinguimos entre, de um lado, a noção de atribuição randômica ao grupo de tratamento dos experimentos e, de outro lado, a amostra aleatória. Consulte o capítulo 4 se precisar relembrar essa diferença.

desejarmos apresentar alguma coisa geral nos baseando em algo específico, teremos algum grau de incerteza. Neste capítulo, discutimos esse processo de inferência estatística, incluindo as ferramentas que cientistas sociais usam para aprender sobre a população pela qual estão interessados por meio da utilização de dados amostrais. Nosso primeiro passo nesse processo é discutir os princípios básicos da teoria probabilística, que, por sua vez, forma a base para toda a inferência estatística.

6.2 NOÇÕES BÁSICAS DE TEORIA PROBABILÍSTICA

Deixe-nos começar com um exemplo.

Suponha que você pegue uma fronha de travesseiro vazia e que, sem que ninguém veja, você meticulosamente separe 550 pequenas bolinhas azuis e 450 pequenas bolinhas vermelhas e as coloque dentro da fronha (totalizando mil bolinhas). Você torce a abertura da fronha algumas vezes para fechá-la e então a chacoalha para misturar as bolinhas. A seguir, você pede para que um amigo coloque a mão dentro da fronha e retire – sem olhar – cem bolinhas e então conte quantas são vermelhas e quantas são azuis.

Obviamente, seu amigo sabe que está retirando apenas uma pequena amostra de bolinhas da população que está dentro da fronha. E, por você ter chacoalhado a fronha e o proibido de olhar dentro da fronha enquanto ele selecionava as cem bolinhas, a seleção representa (mais ou menos) uma amostra aleatória da população. Seu amigo não sabe quantas das bolinhas que estão dentro da fronha são vermelhas e quantas são azuis. Ele apenas sabe quantas bolinhas vermelhas e azuis observou a partir da amostra que retirou da fronha.

Logo após retirar as bolinhas, você pede que ele conte o número de bolinhas azuis e vermelhas. Imaginemos que o resultado seja 46 vermelhas e 54 azuis. Uma vez que ele tenha feito isso, você faz a seguinte pergunta: baseado na sua contagem, qual é o melhor palpite para o percentual de bolinhas vermelhas e de azuis na fronha? O único modo de seu amigo saber com certeza o número é retirando-as da fronha e contando as mil bolinhas. Porém, você não está pedindo ao seu amigo um palpite sem nenhuma informação. Afinal, ele tem alguma informação e pode utilizá-la para formular um palpite melhor do que simplesmente escolher um número entre 0% e 100%.

A partir dos resultados da amostra, ele palpita que 46% das bolinhas que estão dentro da fronha são vermelhas e 54% são azuis. (Reflita um pouco sobre o palpite do seu amigo: embora você saiba que ele está errado, esse é o melhor palpite que ele poderia ter dado considerando o que ele observou, certo?

Antes de informá-lo sobre a resposta correta, você permite que ele recoloque as cem bolinhas de volta à fronha e as misture com as demais bolinhas e pede para que ele repita o processo: ele, novamente, retira cem novas bolinhas e conta o número de vermelhas e azuis. Dessa vez, ele retirou 43 bolinhas vermelhas e 57 azuis.

Você pergunta se ele gostaria de mudar o palpite e, baseado nas novas informações e no cálculo rápido de uma média, ele revisa o palpite e diz que acha que 44,5% das

bolinhas são vermelhas e 55,5% são azuis. (Ele dá esse palpite a partir da média simples de 46% das bolinhas vermelhas da primeira amostra e 43% das bolinhas da segunda amostra.)

As leis da probabilidade são úteis de muitos modos – no cálculo de chances em apostas, por exemplo –, mas, no exemplo acima, elas são úteis pois possibilitam que, a partir de uma determina informação sobre uma característica de uma amostra observada dos dados, possamos tentar generalizar a informação para a população subjacente e não observada. As amostras observadas acima são as duas amostras de cem casos que seu amigo retirou da fronha. A população subjacente é representada pelas mil bolinhas na fronha.

Claramente, o exemplo acima possui limitações. Em especial, no exemplo, você sabe as características reais da população – existem 450 bolinhas vermelhas e 550 bolinhas azuis. Na realidade social, não existe conhecimento comparável do valor das verdadeiras características de uma população subjacente.

Agora passamos a algumas definições.

Um **evento** é o resultado de uma observação aleatória. Dois ou mais eventos podem ser chamados de **eventos independentes** se a realização de um dos eventos não afeta a realização dos demais. Por exemplo, o lançamento de dois dados representa eventos independentes, porque o lançamento do primeiro dado não afeta o resultado do lançamento do segundo.

A probabilidade possui algumas propriedades fundamentais. Primeiro, todos os eventos possuem alguma probabilidade de ocorrer e essa chance varia de 0 a 1. Uma probabilidade de valor 0 significa que o evento é impossível, e uma probabilidade com valor igual a 1 significa que o evento acontecerá com absoluta certeza. Por exemplo, se lançarmos dois dados honestos e somarmos as faces voltadas para cima, a probabilidade da soma das faces ser igual a 13 é 0, uma vez que o valor mais alto possível é 12.

Segundo, a soma de todos os eventos possíveis deve ser exatamente 1. Ou seja, sempre que fazemos uma observação aleatória de um conjunto de eventos possíveis, devemos observar um desses eventos. Por exemplo, se você jogar uma moeda para cima, a probabilidade de o resultado ser cara é de $1/2$, a probabilidade de ser coroa é de $1/2$ $1/2$ e a probabilidade de ser cara ou coroa é de 1, porque $1/2 + 1/2 = 1$.

Terceiro, se (mas somente se!) dois eventos forem independentes, então a probabilidade de esses dois eventos ocorrerem é igual ao produto das chances individuais. Então, se você tem uma moeda não viciada e lançá-la três vezes – tenha em mente que cada lançamento é um evento independente –, a chance de o resultado dos três lançamentos ser igual a coroa é de $1/2 \times 1/2 \times 1/2 = 1/8$.

Obviamente, muitos dos eventos nos quais estamos interessados não são independentes. E, nessas circunstâncias, regras de probabilidade mais complexas, que estão além do escopo dessa discussão, são requeridas.

Por que a probabilidade é relevante para a investigação científica e, em particular, para a ciência política? Por algumas razões. Primeiro, porque cientistas políticos trabalham tipicamente com dados amostrais (e não populacionais), as regras da probabilidade nos dizem como podemos generalizar da nossa amostra para a população

mais ampla. Segundo, e de maneira relacionada, as regras da probabilidade são a chave para identificar quais relações são "estatisticamente significantes" (um conceito que definiremos no próximo capítulo). Colocando de modo diferente, utilizamos a teoria probabilística para decidir se os padrões de relações que observamos em uma amostra podem ter acontecido apenas em razão do acaso.

6.3 APRENDENDO SOBRE A POPULAÇÃO A PARTIR DE UMA AMOSTRA: O TEOREMA DO LIMITE CENTRAL

As razões pelas quais cientistas sociais utilizam dados amostrais em vez de dados populacionais – apesar do fato de nos interessarmos pelos resultados na população em vez de na amostra – são fáceis de entender. Considere uma campanha eleitoral na qual a mídia, o público e os políticos envolvidos querem saber quais são os candidatos preferidos do público e quão preferidos eles são. Em tais circunstâncias, é prático conduzir um censo? Claro que não. A população adulta dos Estados Unidos é de aproximadamente 200 milhões de pessoas; chega ser um eufemismo dizer que não podemos entrevistar cada um desses indivíduos. Simplesmente não temos tempo nem dinheiro para fazer isso. Essa é uma das razões pelas quais o governo dos Estados Unidos conduz um **censo** apenas uma vez a cada dez anos[4].

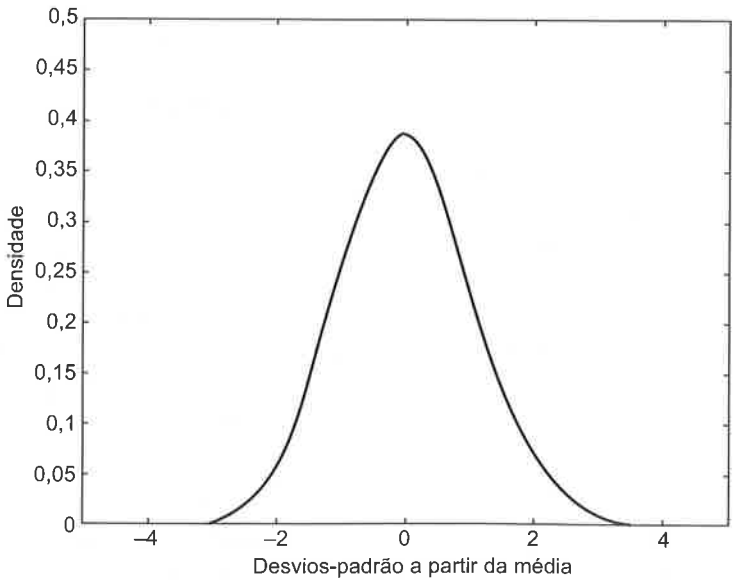

Figura 6.1 – Distribuição de probabilidade normal.

[4] Você pode não estar consciente de que, embora o governo federal conduza somente um censo a cada dez anos, ele conduz *surveys* amostrais com grande frequência na tentativa de mensurar características populacionais, tais como atividade econômica. [No Brasil, o governo federal conduz os Censos Demográficos a cada dez anos e, nos demais anos, realiza a Pesquisa Nacional de Amostra por Domicílio (PNAD) (N.T.).]

Obviamente, todos que estão familiarizados com as pesquisas de opinião sabem que estudiosos e organizações da mídia conduzem rotineiramente *surveys* com amostras da população americana e utilizam os resultados para generalizar sobre a população como um todo. Quando refletimos sobre isso, parece um pouco audacioso imaginar que você pode entrevistar talvez menos que mil pessoas e então utilizar os resultados dessas entrevistas para generalizar sobre as crenças e opiniões de 200 milhões. Como isso é possível?

A resposta recai em um resultado fundamental da estatística chamado **teorema do limite central**, que um estatístico holandês chamado Henk Tijms (2004) define como "o soberano não oficial da teoria probabilística". Antes de mergulharmos na demonstração do teorema e como ele se aplica às pesquisas em ciência sociais, precisamos explorar uma das mais úteis distribuições probabilísticas da estatística, a **distribuição normal**.

6.3.1 A DISTRIBUIÇÃO NORMAL

Dizer que uma distribuição é "normal" *não* significa dizer que ela é "típica" ou "desejável" ou "boa". Uma distribuição que não é "normal" não é algo estranho, como uma distribuição "desviante" ou "anormal". Também é importante enfatizar que distribuições normais não são necessariamente comuns no mundo real. Porém, como veremos, elas são incrivelmente úteis no mundo da estatística.

A distribuição normal é frequentemente chamada de "curva em formato de sino" na linguagem comum. Na Figura 6.1 temos a curva normal e suas diversas propriedades. Primeiro, ela é simétrica em torno da sua média[5], de tal modo que a moda, a mediana e a média são iguais. Segundo, a distribuição normal possui áreas abaixo da curva com distâncias específicas definidas a partir da média. Começando da média e adicionando um desvio-padrão para cada uma das direções, temos uma cobertura de 68% de toda a área abaixo da curva. Adicionando mais um desvio-padrão em cada uma das direções, passamos a ter 95% do total da área[6]. Adicionando um terceiro desvio-padrão em cada direção, temos 99% da área total da curva capturada. Essa característica é comumente conhecida como **regra do 68-95-99** e é exemplificada na Figura 6.2. Você deve ter em mente que essa é uma característica especial da distribuição normal e não se aplica a nenhuma outra forma de distribuição. O que a distribuição normal e a regra do 68-95-99 têm a ver com o processo de aprendizado sobre as características da população a partir de uma amostra?

A distribuição das observações reais de uma amostra – chamada de **distribuição de frequências**, que representa a frequência de cada valor de uma determinada variável – de qualquer variável pode ou não ter o formato da curva normal.

---

[5] De modo equivalente, mas um pouco mais formalmente, podemos caracterizar a distribuição por sua média e variância (ou desvio-padrão) – o que implica que suas obliquidade e curtose são iguais a zero.

[6] Para termos exatamente 95% da área abaixo da curva, precisaríamos de 1,96 desvio-padrão, e não dois desvios-padrão em cada uma das direções a partir da média. Todavia, a regra de dois desvios é considerada uma regra de bolso para muitos dos cálculos estatísticos.

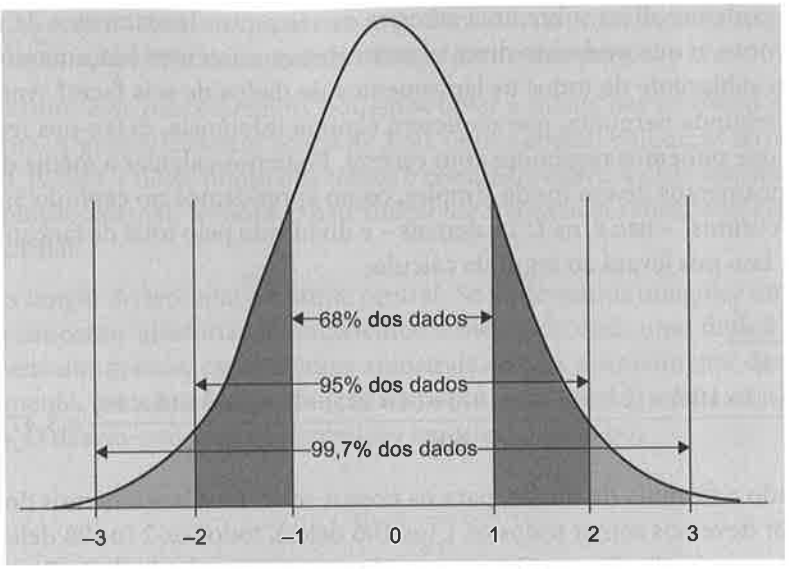

Figura 6.2 – A regra do 68-95-99.

Considere a distribuição de seicentos lançamentos de um dado (honesto) de seis faces, apresentada na Figura 6.3. Note que há algo na figura um pouco estranho: a frequência da distribuição de modo algum se assemelha a uma distribuição normal[7]. Se lançarmos um dado de seis faces honesto seiscentas vezes, quantas vezes devemos observar resultados iguais a 1, 2 etc.? Na média, cem vezes cada, certo? Isto é *bastante próximo* do que observamos na Figura 6.3, mas apenas bastante próximo. Apenas em razão do acaso, temos, por exemplo, um pouco mais de resultados 1 e um pouco menos de resultados 6.

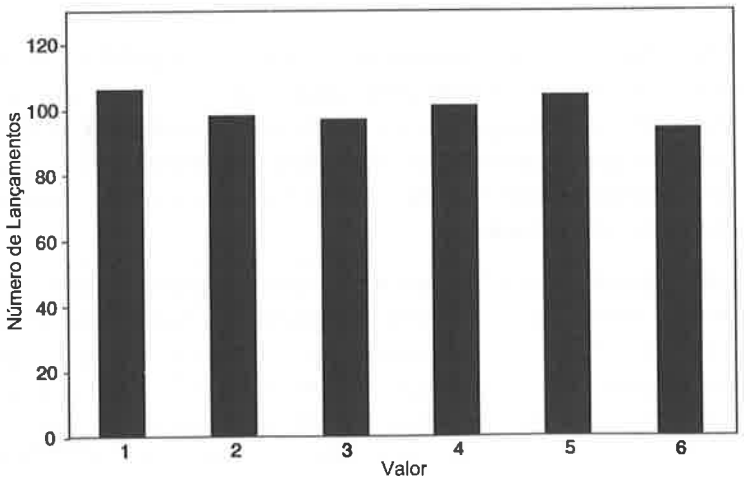

Figura 6.3 – Distribuição de frequência de seiscentos lançamentos de um dado.

[7] De fato, a distribuição é bastante semelhante a uma distribuição uniforme ou achatada.

O que podemos dizer sobre uma amostra de seiscentos lançamentos de um dado? Mais ao ponto, o que podemos dizer, a partir desses seiscentos lançamentos, sobre a população subjacente de todos os lançamentos de dados de seis faces? Antes de responder à segunda pergunta, que requererá alguma inferência, deixe-nos responder à primeira, que podemos responder com certeza. Podemos calcular a média do resultado dos lançamentos de um modo simples, como aprendemos no capítulo 5: somando todos os "eventos" – isto é, os 1, 2 e demais – e dividindo pelo total de lançamentos, no caso, 600. Isso nos levará ao seguinte cálculo:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$
$$= \frac{\sum (1\times 106)+(2\times 98)+(3\times 97)+(4\times 101)+(5\times 104)+(6\times 94)}{600} = 3,47.$$

Seguindo a fórmula da média, para os nossos seiscentos lançamentos do dado, no numerador devemos somar todos os 1 (os 106 deles), todos os 2 (os 98 deles) e assim por diante, e então dividir por 600 para produzir nosso resultado de 3,47.

Podemos calcular o desvio-padrão dessa distribuição pela fórmula:

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{1753,40}{599}} = 1,71.$$

Observando o numerador da fórmula do desvio-padrão que aprendemos no capítulo 5, observamos que $\sum(Y_i - \bar{Y})^2$ indica que, para cada observação (1, 2, 3, 4, 5 ou 6), subtraímos o valor da média (3,47), então elevamos ao quadrado a diferença, somamos todas as 600 diferenças da média elevadas ao quadrado, que produz o numerador 1.753,40 abaixo do sinal de raiz quadrada. Dividimos esse valor por 599 (isto é, $n-1$), então tiramos a raiz quadrada e temos como resultado o desvio-padrão de 1,71.

Como pontuamos, a média da amostra é de 3,47, mas qual deveria ser a média esperada? Se realizássemos exatamente cem lançamentos com cada um dos lados do dado como resultado, a média esperada seria 3,50, então a média de nossa amostra é um pouco menor do que a que esperaríamos obter. Novamente, podemos observar que nossos lançamentos possuem "muitos" 1 e uns "poucos" 6, assim a nossa média ser um pouco menor que 3,50 faz sentido.

O que aconteceria, contudo, se lançássemos o mesmo dado outras seiscentas vezes? Qual seria a média obtida a partir desses lançamentos? Obviamente, não podemos afirmar com certeza. Talvez conseguíssemos outra amostra com média de 3,47, ou talvez ela fosse um pouco maior que 3,50, ou talvez a média fosse exatamente 3,50, Suponha que lançamos o dado seiscentas vezes, não uma, nem duas, mas um número infinito de vezes. Sejamos claros: *não estamos dizendo para supor um número infinito de lançamento dos dados*, mas que os *seiscentos lançamentos sejam repetidos infinitas vezes*. Essa é uma distinção crítica. Imaginamos que estamos coletando uma amostra de seiscentos lançamentos, não uma, mas um número infinito de vezes. Podemos chamar essa hipotética distribuição das médias amostrais de **distribuição amostral.** Ela

é hipotética porque cientistas quase nunca podem coletar mais de uma amostra para uma população subjacente em um determinado ponto do tempo.

Se seguirmos esse procedimento, podemos obter a média das amostras e as expor graficamente. Algumas estariam acima de 3,50, outras abaixo e algumas seriam exatamente 3,50. Porém, é nesse ponto que temos o resultado-chave: a distribuição amostral terá distribuição normal, embora a distribuição de frequência subjacente, claramente, não seja normal.

Esse é o *insight* do teorema do limite central. Se pudéssemos imaginar um número infinito de amostras aleatórias e plotássemos a média de cada uma dessas amostras aleatórias em um gráfico, essas médias amostrais seriam normalmente distribuídas. Adicionalmente, a média da distribuição amostral seria igual à média da verdadeira população. O desvio-padrão da distribuição amostral é dado por

$$\sigma_{\bar{Y}} = \frac{s_Y}{\sqrt{n}},$$

em que n é o tamanho da amostra. O desvio-padrão da distribuição amostral da média das amostras, que é conhecido como **erro-padrão da média** (ou "erro-padrão"), é simplesmente igual ao desvio-padrão da amostra dividido pela raiz quadrada do tamanho da amostra. No nosso exemplo anterior do lançamento de dados, nosso erro-padrão da média é dado por

$$\sigma_{\bar{Y}} = \frac{1,71}{\sqrt{600}} = 0,07.$$

Lembre-se que nosso objetivo aqui é aprender sobre a população subjacente utilizando o que temos certeza que sabemos sobre a amostra. Sabemos que a média da nossa amostra de seiscentos lançamentos é 3,47 e seu desvio-padrão de 1,71. Dessas características, podemos imaginar que, se fizermos infinitas vezes o lançamento do dado seiscentas vezes, o resultado da distribuição amostral teria um desvio-padrão de 0,07. Nossa melhor aproximação da média da população é 3,47, porque esse é o resultado gerado pela nossa amostra[8]. Mas sabemos que nossa amostra de seiscentos lançamentos pode ser ligeiramente diferente da média verdadeira da população, podendo ser tanto um pouco maior quanto um pouco menor. O que podemos fazer, portanto, é usar nosso conhecimento de que a distribuição amostral é normal e invocar a regra do 68-95-99 para criar um **intervalo de confiança** sobre a provável localização da média da população.

Como fazemos isso? Primeiro, escolhemos o grau de confiança que queremos ter em nossa estimativa. Embora possamos escolher qualquer grau de confiança entre 0 e 100, cientistas sociais tradicionalmente escolhem o intervalo de confiança 95%. Se seguirmos esse padrão – e por nossa distribuição amostral ser normal –, começaría-

[8] Alguns podem imaginar que nosso melhor palpite deveria ser 3,50, porque, em teoria, um dado honesto deve produzir tal resultado.

mos com nossa média (3,47) e nos moveríamos dois erros-padrão da média em cada uma das direções para produzir o intervalo no qual, com 95% de confiança, se encontra a média da população. Por que dois erros-padrão? Porque com dois erros-padrão temos uma área de 95% abaixo da curva. Novamente, para termos precisamente 95% de confiança abaixo da curva, moveríamos 1,96 e não dois erros-padrão em cada direção. Mas a regra de bolso de usar dois erros-padrão é uma prática comum. Em outras palavras,

$$\bar{Y} \pm 2 \times \sigma_{\bar{Y}} = 3,47 \pm (2 \times 0,07) = 3,47 \pm 0,14 .$$

Isso significa, para nossa amostra, que temos 95% de confiança de que a média da população de nosso lançamento de dados está em algum lugar no intervalo entre 3,33 e 3,61.

É possível que estejamos errados e que a média da população esteja fora do intervalo? Sim, e ainda sabemos *quão* provável é que ela esteja fora do intervalo. Existem 2,5% de chance de que a média da população seja menor que 3,33 e 2,5% de chance de que a média da população seja maior que 3,61, em um total de 5% de chance de que a média da população não esteja no intervalo de 3,33 a 3,61. Por diversas razões, é possível que queiramos ter mais confiança em nossa estimação. Digamos que, em vez de termos 95% de confiança, gostaríamos de ter 99% de confiança em nossas estimativas. Nesse caso, simplesmente nos moveríamos *três* (em vez de dois) erros-padrão em cada uma das direções a partir da média da nossa amostra (3,47), gerando um intervalo de 3,26-3,68.

Ao longo deste exemplo temos sido ajudados pelo fato de sabermos as características subjacentes ao processo de geração dos dados (um dado honesto). No mundo real, cientistas sociais quase nunca possuem essa vantagem. Na próxima seção, apresentamos um exemplo em que não a temos.

6.4 EXEMPLO: TAXAS DE APROVAÇÃO PRESIDENCIAL

Entre 20 e 24 de junho de 2012, a NBC News e o *Wall Street Journal* promoveram um *survey* em que mil americanos foram selecionados aleatoriamente para responder a perguntas sobre suas crenças políticas. Entre as questões, estava uma que pretendia capturar a opinião do entrevistado sobre o desempenho do presidente em exercício:

De maneira geral, você aprova ou desaprova o trabalho que Barack Obama está fazendo como presidente?

Essa é uma pergunta amplamente conhecida e utilizada há mais de meio século por quase todas as organizações de pesquisa[9]. Em junho de 2012, 47% da amostra aprovava o trabalho que Obama estava fazendo, 48% desaprovavam e 5% não tinham certeza sobre como avaliar o trabalho do presidente[10].

---

[9] Evidentemente, a única diferença é o nome do presidente em exercício.

[10] A fonte do *survey* é: <http://www.pollingreport.com/obama_job2.htm>. Acesso em: 11 jul. 2012.

Essas organizações midiáticas, claramente, não estão interessadas na opinião desses mil americanos que compõem a amostra, exceto na medida em que ela diz algo sobre a população adulta como um todo. É possível utilizar as respostas desses mil entrevistados para fazer exatamente isso utilizando a lógica do teorema do limite central e as ferramentas previamente descritas.

Para reiterar, temos certeza de que conhecemos as propriedades da nossa amostra aleatória de mil pessoas. Se considerarmos que as 470 respostas de aprovação são equivalentes a 1 e as 530 respostas remanescentes são equivalentes a 0, então podemos calcular a média da nossa amostra, $\bar{Y}$, por[11]:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum(470 \times 1) + (530 \times 0)}{1.000} = 0,47 .$$

Calculamos o desvio-padrão da média, s_Y, do seguinte modo:

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{470(1-0,47)^2 + 530(0-0,47)^2}{1.000-1}}$$
$$= \sqrt{\frac{249,1}{999}} = 0,50.$$

Mas o que podemos dizer sobre a população como um todo? Obviamente, à diferença da média da amostra, não temos certeza sobre o valor da média da população. Mas se imaginarmos que, em vez de uma amostra de mil respondentes, temos um infinito número de amostras de mil pessoas, então o teorema do limite central nos diz que as médias dessas amostras seriam distribuídas normalmente. Nosso melhor palpite para a média da população, claro, é 0,47, porque essa é a média de nossa amostra. O erro-padrão da média é dado por

$$\sigma_{\bar{Y}} = \frac{0,50}{\sqrt{1.000}} = 0,016,$$

que consiste em uma medida de incerteza sobre a média da população. Se utilizarmos nossa regra de bolso e calcularmos o intervalo de 95% de confiança adicionando dois erros-padrão para cada lado a partir do valor da média da amostra, teremos o seguinte intervalo:

$$\bar{Y} \pm 2 \times \sigma_{\bar{Y}} = 0,47 \pm (2 \times 0,016) = 0,47 \pm 0,032 ,$$

ou entre 0,438 e 0,502, que traduz em um intervalo de 95% de confiança que o valor da média da aprovação de Obama na população está entre 43,8% e 50,2%.

[11] Existem diferentes modos para lidar matematicamente com os 5% de respostas "sem certeza". Nesse caso, como estamos interessados em calcular a taxa de "aprovação", é razoável juntar as respostas de desaprovação e de incerteza. É muito importante comunicar exatamente o que estamos fazendo à nossa audiência quando tomamos uma decisão desse tipo, para que nosso trabalho possa ser avaliado de maneira correta.

É daqui que os sinais de "±" que sempre vemos nas pesquisas de opinião surgem[12]. O melhor palpite da média da população é o valor da média da amostra mais ou menos dois erros-padrão. Então os dados que costumamos observar com os sinais de "±" são, normalmente, construídos com intervalos de 95% de confiança.

6.4.1 QUE TIPO DE AMOSTRA ERA?

Se você ler o exemplo anterior com atenção, notará que descrevemos a pesquisa da NBC-*Wall Street Journal* como uma amostra aleatória de mil indivíduos. Isso significa que eles usaram algum mecanismo (como a discagem aleatória de número de telefone) para assegurar que todos os membros da população tivessem igual probabilidade de serem selecionados para o *survey*. É preciso reiterar a importância de usar amostras aleatórias. O teorema do limite central *somente* se aplica a amostras selecionadas aleatoriamente. Com amostras de conveniência não podemos invocar o teorema do limite central para construir uma distribuição amostral e criar um intervalo de confiança.

Esta lição é crucial: uma amostra de conveniência selecionada de modo não aleatório pouco ajuda na tarefa de estabelecer um elo entre a amostra e a população que queremos estudar. Isso produz vários tipos de implicações das "pesquisas" que organizações midiáticas realizam em seus *sites*. O que tais "*surveys*" dizem sobre a população como um todo? Como a amostra deles claramente não é uma seleção aleatória da população subjacente, a resposta é "nada".

Existe uma lição relacionada a esta. O exemplo anterior representa uma conexão direta entre uma amostra (as mil pessoas do *survey*) e a população (todos os adultos nos Estados Unidos). Frequentemente a conexão entre a amostra e a população é menos direta. Considere, por exemplo, o exame das votações nominais em um legislativo em um determinado ano. Assumindo que seja bastante fácil coletar todas as votações nominais para cada um dos parlamentares (que consiste em nossa amostra), somos confrontados com uma pergunta ligeiramente desconcertante: qual é a nossa população de interesse? A resposta não é óbvia e nem todos os cientistas sociais concordariam com ela. Alguns podem dizer que os dados não representam uma amostra, mas uma população, porque o banco de dados contém os votos de todos os parlamentares. Outros podem argumentar que ela é uma amostra de um ano de trabalho no legislativo desse a sua criação. Outros, ainda, poderiam dizer que a amostra é um dos eventos possíveis de um infinito número de legislativos que poderiam ser observados nesse ano em particular. É suficiente dizer que, neste exemplo, não existe um consenso científico claro do que constituiria a "amostra" e o que constituiria a "população".

6.4.2 UMA NOTA SOBRE OS EFEITOS DO TAMANHO DA AMOSTRA

Como a fórmula do intervalo de confiança indica, quanto menor o erro-padrão, mais "estreito" nosso intervalo de confiança será; quanto maior o erro-padrão, mais

[12] Na prática, a maioria das empresas de pesquisa possuem seus próprios ajustes adicionais para realizar esses cálculos, mas eles começam sempre com essa lógica básica.

"alargado" nosso intervalo de confiança será. Se estivermos interessados em estimar valores populacionais, nos baseando em nossas amostras, com a maior precisão possível, então é desejável ter um intervalo de confiança mais estreito do que alargado.

Como podemos conseguir isso? Pela fórmula do erro-padrão da média fica claro, utilizando álgebra simples, que podemos obter valores menores para o erro-padrão de duas formas: com um numerador menor ou um denominador maior. Como obter um numerador menor – o desvio-padrão da amostra – não é algo que podemos fazer na prática, podemos considerar se é possível ter um denominador maior – isto é, aumentar o tamanho da amostra.

Amostras grandes reduzirão o tamanho dos erros-padrão e amostras menores aumentarão o tamanho dos erros-padrão. Esperamos que isso seja intuitivo para você. Se tivermos uma amostra grande, então deve ser mais fácil realizar inferências sobre a população de interesse; amostras menores, por sua vez, devem produzir menos confiança sobre a estimativa populacional.

Em nosso exemplo anterior, se em vez de uma amostra de mil pessoas tivéssemos uma amostra muito maior – digamos 2.500 pessoas –, nossos erros-padrão seriam dados por

$$\sigma_{\bar{Y}} = \frac{0,50}{\sqrt{2.500}} = 0,010,$$

que consiste em menos de dois terços do tamanho do nosso erro-padrão real de 0,016. Você pode ver matematicamente que adicionar dois erros-padrão de 0,010 em cada uma das direções a partir da média produz um intervalo de confiança mais estreito do que dois erros-padrão de 0,016. Mas note que o custo da redução de 1,2% em cada uma das direções do nosso intervalo de confiança é de quase mais 1.500 respondentes e que, em muitos casos, essa redução do erro não compensará nem pelo tempo, nem pelos custos financeiros envolvidos na obtenção de entrevistas adicionais.

Considere o caso oposto. Se, em vez de mil entrevistados, nossa amostra fosse composta por apenas quatrocentos, nosso erro-padrão seria:

$$\sigma_{\bar{Y}} = \frac{0,50}{\sqrt{400}} = 0,025\,,$$

que, quando multiplicado por dois para produzir o intervalo de confiança de 95%, nos levaria a uma soma 0,05 (ou 5%) em cada direção.

Poderíamos ser completamente tolos e selecionar uma amostra aleatória de apenas 64 pessoas se quiséssemos. Isso geraria um intervalo de confiança bastante alargado. O erro-padrão seria

$$\sigma_{\bar{Y}} = \frac{0,50}{\sqrt{64}} = 0,062\,,$$

que, quando multiplicado por dois para produzir o intervalo de confiança de 95%, nos levaria a uma soma de robusto 0,124 (ou 12,4%) em cada direção. Nessas circuns-

tâncias, diríamos que a aprovação de Obama pela população era de 47%, mas, com o intervalo de confiança de 95%, poderia estar entre 34,6% e 59,4%. Esse intervalo seria muito amplo para informar qualquer coisa sobre a aprovação do presidente.

Em suma, a resposta para a pergunta: "Quão grande minha amostra precisa ser?" é outra pergunta: "Quão estreito você quer que seu intervalo de confiança seja?".

6.5 OLHANDO ADIANTE: EXAMINANDO RELAÇÕES ENTRE VARIÁVEIS

Façamos um balanço por um momento. Neste livro, temos enfatizado que a pesquisa em ciência política envolve a avaliação de explicações causais, o que implica o exame de relações entre duas ou mais variáveis. Porém, neste capítulo, tudo que fizemos foi falar sobre o processo de inferência estatística com uma única variável. Esse foi um desvio necessário, porque tínhamos que ensinar a você a lógica da inferência estatística – isto é, como utilizamos amostras para aprender algo sobre uma população subjacente.

No capítulo 7, você aprenderá três diferentes modos de fazer testes empíricos bivariados. Examinaremos relações entre duas variáveis, tipicamente em uma amostra, e então faremos avaliações sobre a chance dessas relações existirem na população. A lógica é idêntica à que você acabou de aprender; meramente estendemos o que apresentamos neste capítulo para a análise da relação entre duas variáveis. Por fim, no capítulo 8, você aprenderá outro modo de conduzir testes de hipóteses envolvendo duas variáveis – o modelo de regressão bivariado.

CONCEITOS INTRODUZIDOS NESTE CAPÍTULO

- Amostra – um subconjunto de casos de uma população de interesse.
- Amostra aleatória – uma amostra em que cada um dos membros da população subjacente possui igual chance de ser selecionado.
- Censo – uma pesquisa com toda a população.
- Distribuição amostral – distribuição hipotética das médias das amostras.
- Distribuição de frequência – uma distribuição dos valores reais de uma amostra.
- Distribuição normal – uma distribuição estatística com forma de sino que pode ser caracterizada inteiramente pela sua média e seu desvio-padrão.
- Erro-padrão da média – o desvio-padrão da distribuição amostral das médias das amostras.
- Evento – um resultado de uma observação aleatória.
- Eventos independentes – dois ou mais eventos em que a realização de um evento não afeta a realização do outro.

- Inferência estatística – o processo de utilizar o que sabemos sobre uma amostra para fazer afirmações probabilísticas sobre uma população mais ampla.
- Intervalo de confiança – uma afirmação probabilística sobre quão provável é o valor de uma característica da população baseada nas observações de uma amostra.
- População – dados para todos os possíveis casos relevantes.
- Regra do 68-95-99 – característica útil da distribuição normal que estabelece que, ao nos movermos 1, 2 e 3 desvios-padrão em cada direção a partir da média, teremos uma cobertura de 68%, 95% e 99% da área abaixo da curva normal.
- Teorema do limite central – resultado fundamental da estatística que indica que, se alguém coletasse um número infinito de amostras aleatórias e plotasse os resultados das médias das amostras, essas médias seriam normalmente distribuídas ao redor da verdadeira média da população.

EXERCÍCIOS

1. Vá ao *site* http://www.pollingreport.com e encontre uma estatística que interesse a você. Tenha a certeza de clicar na opção *full details*, quando disponível, para saber o tamanho da amostra que respondeu a esse item no *survey*. Então calcule os intervalos de confiança de 95% e 99% do valor populacional da estatística que você encontrou. Mostre todos os seus cálculos, imprima a página do *site* e entregue-a junto com o exercício[13].
2. Considerando o mesmo item do *survey* que você escolheu na pergunta anterior, o que aconteceria com o intervalo de confiança se o tamanho da amostra fosse reduzido pela metade? E o que aconteceria se ele dobrasse? Suponha que o desvio-padrão da amostra não se altera. Mostre todos os cálculos.
3. Amostras maiores são sempre melhores do que amostras menores? Explique sua resposta.
4. Volte à Tabela 5.2, que mostra o percentual de voto do partido do incumbente nas eleições presidenciais nos Estados Unidos. Calcule o erro-padrão da média para essa distribuição e, então, construa um intervalo de confiança de 95% para a média da população. Mostre todos os cálculos. O que o intervalo de confiança de 95% nos diz sobre esse caso específico?
5. Se obtivermos uma amostra representativa da população dos EUA de mil entrevistados para uma pergunta específica de um *survey* e obtivermos uma margem de confiança de 95%, quantos entrevistados seriam necessários para você obter o mesmo intervalo para a população de Maine, assumindo que a distribuição da resposta é a mesma para ambas as populações?

[13] Para dados de pesquisas de opinião pública no Brasil, conferir: <http://pollingdata.com.br/ ou http://www.ibope.com.br/pt-br/conhecimento/relatoriospesquisas/Paginas/default.aspx>. [N.T.]