Capítulo 10

# Introdução à Inferência Estatística

## 10.1 Introdução

Vimos, na Parte 1, como resumir descritivamente variáveis associadas a um ou mais conjuntos de dados. Na Parte 2, construímos modelos teóricos (probabilísticos), identificados por parâmetros, capazes de representar adequadamente o comportamento de algumas variáveis. Nesta terceira parte apresentaremos os argumentos estatísticos para fazer afirmações sobre as características de uma população, com base em informações dadas por amostras.

O uso de informações de uma amostra para concluir sobre o todo faz parte da atividade diária da maioria das pessoas. Basta observar como uma cozinheira verifica se o prato que ela está preparando tem ou não a quantidade adequada de sal. Ou, ainda, quando um comprador, após experimentar um pedaço de laranja numa banca de feira, decide se vai comprar ou não as laranjas. Essas são decisões baseadas em procedimentos amostrais.

Nosso objetivo nos capítulos seguintes é procurar dar a conceituação formal a esses princípios intuitivos do dia-a-dia para que possam ser utilizados cientificamente em situações mais complexas.

## 10.2 População e Amostra

Nos capítulos anteriores, tomamos conhecimento de alguns modelos probabilísticos que procuram medir a variabilidade de fenômenos casuais de acordo com suas ocorrências: as distribuições de probabilidades de variáveis aleatórias (qualitativas ou quantitativas). Na prática, freqüentemente o pesquisador tem alguma idéia sobre a forma da distribuição, mas não dos valores exatos dos parâmetros que a especificam.

Por exemplo, parece razoável supor que a distribuição das alturas dos brasileiros adultos possa ser representada por um modelo normal (embora as alturas não possam assumir valores negativos). Mas essa afirmação não é suficiente para determinar qual a distribuição normal correspondente; precisaríamos conhecer os parâmetros (média e variância) dessa normal para que ela ficasse completamente especificada. O propósito do pesquisador seria, então, descobrir (estimar) os parâmetros da distribuição para sua posterior utilização.

Se pudéssemos medir as alturas de todos os brasileiros adultos, teríamos meios de obter sua distribuição exata e, daí, produzir os correspondentes parâmetros. Mas nessa situação não teríamos necessidade de usar a inferência estatística!

Raramente se consegue obter a distribuição exata de alguma variável, ou porque isso é muito dispendioso, ou muito demorado ou às vezes porque consiste num processo destrutivo. Por exemplo, se estivéssemos observando a durabilidade de lâmpadas e testássemos todas até queimarem, não restaria nenhuma para ser vendida. Assim, a solução é selecionar parte dos elementos (amostra), analisá-la e *inferir* propriedades para o todo (população).

Outras vezes estamos interessados em explorar relações entre variáveis envolvendo experimentos mais complexos, para a obtenção dos dados. Por exemplo, gostaríamos de obter resposta para a seguinte indagação: a altura que um produto é colocado na gôndola de um supermercado afeta a sua venda? Observe que para responder a questão precisamos obter dados de vendas com o produto oferecido em diferentes alturas, e que essas vendas sejam controladas para evitar interferências de outros fatores que não a altura. Nesse caso não existe claramente um conjunto de *todos* os elementos para os quais pudéssemos encontrar os parâmetros populacionais. Recorrer a modelos para descrever o todo (população) facilita a identificação e solução do problema. Nesse exemplo, supondo que as vendas $V_h$ do produto oferecido na altura $h$ ($h = 1$ representando *baixo*, $h = 2$ representando *meio* e $h = 3$ representando *alto*) segue uma distribuição próxima a normal, ou seja, $V_h \sim N(\mu_h, \sigma^2)$, o nosso problema passa a ser o de verificar, por meio de dados coletados do experimento (amostra), se existe evidência de igualdade das médias $\mu_1$, $\mu_2$ e $\mu_3$. Note que, em nossa formulação do problema, supusemos que as três situações de alturas resultam observações com a mesma variância $\sigma^2$. Essa suposição poderia ser modificada.

Soluções de questões como as apresentadas acima são o objeto da *inferência estatística*.

Dois conceitos básicos são, portanto, necessários para o desenvolvimento da Inferência Estatística: população e amostra.

**Definição.** *População* é o conjunto de todos os elementos ou resultados sob investigação. *Amostra* é qualquer subconjunto da população.

Vejamos outros exemplos para melhor entender essas definições.

**Exemplo 10.1.** Consideremos uma pesquisa para estudar os salários dos 500 funcionários da Companhia MB. Seleciona-se uma amostra de 36 indivíduos, e anotam-se os seus salários. A variável aleatória a ser observada é "salário". A população é formada pelos 500 funcionários da companhia. A amostra é constituída pelos 36 indivíduos selecionados. Na realidade, estamos interessados nos salários, portanto, para sermos mais precisos, devemos considerar como a população os 500 salários correspondentes aos 500 funcionários. Conseqüentemente, a amostra será formada pelos 36 salários dos indivíduos selecionados. Podemos estudar a distribuição dos

salários na amostra, e esperamos que esta reflita a distribuição de todos os salários, desde que a amostra tenha sido escolhida com cuidado.

**Exemplo 10.2.** Queremos estudar a proporção de indivíduos na cidade $A$ que são favoráveis a certo projeto governamental. Uma amostra de 200 pessoas é sorteada, e a opinião de cada uma é registrada como sendo a favor ou contra o projeto. A população consiste de todos os moradores da cidade, e a amostra é formada pelas 200 pessoas selecionadas. Podemos, como foi visto no Capítulo 5, definir a variável $X$, que toma o valor 1, se a resposta de um morador for favorável, e o valor 0, se a resposta for contrária ao projeto. Assim, nossa população pode ser reduzida à distribuição de $X$, e a amostra será constituída de uma seqüência de 200 zeros e uns.

**Exemplo 10.3.** O interesse é investigar a duração de vida de um novo tipo de lâmpada, pois acreditamos que ela tenha uma duração maior do que as fabricadas atualmente. Então, 100 lâmpadas do novo tipo são deixadas acesas até queimarem. A duração em horas de cada lâmpada é registrada. Aqui, a variável é a duração em horas de cada lâmpada. A população é formada por todas as lâmpadas fabricadas ou que venham a ser fabricadas por essa empresa, com o mesmo processo. A amostra é formada pelas 100 lâmpadas selecionadas. Note-se que nesse caso não podemos observar a população, ou seja, a distribuição da duração de vida das lâmpadas na população, pois isso corresponderia a queimar todas as lâmpadas. Assim, em alguns casos, não podemos observar a população toda, pois isso significaria danificar (ou destruir) todos os elementos da população. Esse problema geralmente é contornado atribuindo-se um modelo teórico para a distribuição da variável populacional.

**Exemplo 10.4.** Em alguns casos, fazemos suposições mais precisas sobre a população (ou sobre a variável definida para os elementos da população). Digamos que $X$ represente o peso real de pacotes de café, enchidos automaticamente por uma máquina. Sabe-se que a distribuição de $X$ pode ser representada por uma normal, com parâmetros $\mu$ e $\sigma^2$ desconhecidos. Sorteamos 100 pacotes e medimos seus pesos. A população será o conjunto de todos os pacotes enchidos ou que virão a ser enchidos pela máquina, e que pode ser suposta como normal. A amostra será formada pelas 100 medidas obtidas dos pacotes selecionados, que pode ser pensada como constituída de 100 observações feitas de uma distribuição normal. Veremos mais adiante como tal amostra pode ser obtida.

**Exemplo 10.5.** Para investigar a "honestidade" de uma moeda, nós a lançamos 50 vezes e contamos o número de caras observadas. A população, como no caso do Exemplo 10.2, pode ser considerada como tendo a distribuição da variável $X$, assumindo o valor 1, com probabilidade $p$, se ocorrer cara, e assumindo o valor 0, com probabilidade $1 - p$, se ocorrer coroa. Ou seja, a população pode ser considerada como tendo distribuição de Bernoulli com parâmetro $p$. A variável ficará completamente especificada quando conhecermos $p$. A amostra será uma seqüência de 50 números zeros ou uns.

**Exemplo 10.6.** Há razões para supor que o tempo $Y$ de reação a certo estímulo visual depende da idade do indivíduo (esse exemplo será usado nos Capítulos 15 e 16). Suponha, ainda, que essa *dependência seja linear*. Para verificarmos se essa suposição é verdadeira, obtiveram-se 20 dados da seguinte maneira: 20 pessoas foram selecionadas, sendo 10 homens e 10 mulheres. Dentro de cada grupo de homens e mulheres foram selecionadas duas pessoas das seguintes faixas de idade: 20, 25, 30, 35 e 40 anos. Cada pessoa foi submetida ao teste e seu tempo de reação $y$ foi medido. A população poderia ser considerada como formada por todas aquelas pessoas que viessem a ser submetidas ao teste, segundo o sexo e a idade. A amostra é formada pelas 20 medidas, que estão apresentadas na Tabela 15.1.

**Observações:**

(i) Os três últimos exemplos mostram uma ampliação do conceito definido de população, ou seja, designamos agora a população como sendo a função probabilidade ou função densidade de probabilidade de uma v.a. $X$, modelando a característica de interesse. Esse artifício simplifica substancialmente o problema estatístico, exigindo no entanto uma proposta de modelo para a variável $X$. Nesses casos simplificaremos a linguagem, dizendo: "seja a população $f(x)$". Por exemplo, "considere a população das alturas $X \sim N(\mu, \sigma^2)$".

(ii) Essa abordagem, por meio da distribuição de probabilidades, utiliza muitas vezes o conceito de população infinita contínua, exigindo um tratamento matemático mais cuidadoso. É mais fácil apresentar os problemas e soluções por meio de populações finitas. É o que faremos muitas vezes. Entretanto, é importante que o estudante aprenda a trabalhar com o conceito de modelo, explorando o caso de "população $f(x)$".

## 10.3 Problemas de Inferência

Como já dissemos anteriormente, o objetivo da Inferência Estatística é produzir afirmações sobre dada característica da população, na qual estamos interessados, a partir de informações colhidas de uma parte dessa população. Essa característica na população pode ser representada por uma variável aleatória. Se tivéssemos informação completa sobre a função de probabilidade, no caso discreto, ou sobre a função densidade de probabilidade, no caso contínuo, da variável em questão, não teríamos necessidade de escolher uma amostra. Toda a informação desejada seria obtida por meio da distribuição da variável, usando-se a teoria estudada anteriormente.

Mas isso raramente acontece. Ou não temos qualquer informação a respeito da variável, ou ela é apenas parcial. Podemos admitir, como no exemplo das alturas de brasileiros adultos, que ela siga uma distribuição normal, mas desconhecemos os parâmetros que a caracterizam (média, variância). Em outros casos, podemos ter uma idéia desses parâmetros, mas desconhecemos a forma da curva. Ou ainda, o que é muito freqüente, não possuímos informações nem sobre os parâmetros, nem sobre a forma da curva. Em todos os casos, o uso de uma amostra nos ajudaria a formar uma opinião sobre o comportamento da variável (população).

Embora a identificação e a descrição da população sejam fundamentais no processo inferencial, é comum os pesquisadores dedicarem mais atenção em descrever a amostra do que a população para a qual serão feitas as afirmações. É imprescindível que se explicite claramente a população investigada.

Neste livro estaremos mais preocupados em trabalhar com populações descritas por modelos do que com populações finitas identificadas por elementos portadores de uma característica de interesse. Portanto, na maioria das vezes, iremos nos referir à "população $X$", significando que a variável de interesse $X$, definida sobre a população-alvo, segue uma distribuição $f(x)$. Nosso problema de interesse passaria a ser o de fazer afirmações sobre a forma da curva e seus parâmetros.

Alguns exemplos simples nos darão uma noção dos tipos de formulações e problemas que a inferência estatística pode nos ajudar a resolver.

**Exemplo 10.5.** (continuação) Voltemos ao exemplo da moeda. Indicando por $X$ o número de caras obtidas depois de lançar a moeda 50 vezes, sabemos que, se tomados alguns cuidados quando do lançamento, $X$ segue uma distribuição binomial, ou seja, $X \sim b(50, p)$. Esse modelo é válido, admitindo-se ou não a "honestidade" da moeda, isto é, sendo ou não $p = 1/2$. Lançada a moeda, vamos supor que tenham ocorrido 36 caras. Esse resultado traz evidência de que a moeda seja "honesta"? Para tomarmos uma decisão, podemos partir do princípio de que a moeda não favorece nem cara nem coroa, isto é, $p = 1/2$. Com essa informação e com o modelo binomial, podemos encontrar qual a probabilidade de se obterem 36 caras ou mais, e esse resultado nos ajudaria a tomar uma decisão. Suponha que a decisão foi rejeitar a "honestidade" da moeda: qual é a melhor estimativa para $p$, baseando-se no resultado observado?

Descrevemos aí os dois problemas básicos da Inferência Estatística: o primeiro é chamado *teste de hipóteses*, e o segundo, *estimação*. Nos capítulos seguintes, esses problemas serão abordados com mais detalhes.

**Exemplo 10.4.** (continuação) Às vezes, o modelo teórico associado ao problema não é tão evidente. No caso da máquina de encher pacotes de café automaticamente, digamos que ela esteja regulada para enchê-los segundo uma distribuição normal com média 500 gramas e desvio padrão de 100 gramas, isto é, $X \sim N(500, 20^2)$. Sabemos também que, às vezes, a máquina desregula-se e, quando isso acontece, o único parâmetro que se altera é a média, permanecendo a mesma variância. Para manter a produção sob controle, iremos colher uma amostra de 100 pacotes e pesá-los. Como essa amostra nos ajudará a tomar uma decisão? Parece razoável, nesse caso, usarmos a média $\bar{x}$ da amostra como informação pertinente para uma decisão. Mesmo que a máquina esteja regulada, dificilmente $\bar{x}$ será igual a 500 gramas, dado que os pacotes apresentam certa variabilidade no peso. Mas se $\bar{x}$ não se afastar muito de 500 gramas, não existirão razões para suspeitarmos da qualidade do procedimento de produção. Só iremos pedir uma revisão se $\bar{x} - 500$, em valor absoluto, for "muito grande".

O problema que se apresenta agora é o de decidir o que é próximo ou distante de 500 gramas. Se o mesmo procedimento de colher a amostra de 100 pacotes fosse repetido um número muito grande de vezes, sob a condição de a máquina estar regulada, teríamos idéia do comportamento da v.a. $\bar{x}$ , e saberíamos dizer se aquele valor observado é ou não um evento raro de ocorrer. Caso o seja, é mais fácil suspeitar da regulagem da máquina do que do acaso.

Vemos, então, a importância nesse caso de se conhecer as propriedades da distribuição da variável $\bar{x}$.

**Exemplo 10.6.** (continuação) A descrição matemática da v.a. $Y$: tempo de reação ao estímulo é um pouco mais complexa. Podemos supor que esse tempo, para uma dada idade $x$, seja uma v.a. com distribuição normal, com média dependendo da idade $x$, ou seja, podemos escrever

$$Y \sim N(\mu(x),\ \sigma^2).$$

A *linearidade* expressa no problema pode ser incluída na média $\mu(x)$ da seguinte maneira:

$$\mu(x) = \alpha + \beta x.$$

Voltaremos a esse modelo no Capítulo 16. Outra maneira de escrever as duas relações anteriores é

$$Y|\, x \sim N(\alpha + \beta x;\ \sigma^2).$$

Leia-se "$Y$ dado $x$".

Podemos, por exemplo, estimar os parâmetros $\alpha$ e $\beta$, baseados na amostra de 20 dados. Ou podemos querer investigar a possibilidade de $\beta$ ser igual a zero, significando que a idade não afeta o tempo de reação. Novamente, os dois principais problemas de inferência aparecem aqui: estimação e teste de uma hipótese. Um outro problema importante em inferência é o de *previsão*. Por exemplo, considerando um grupo de pessoas de 40 anos, poderemos prever com o modelo acima qual será o respectivo tempo de reação.

Repetir um mesmo experimento muitas vezes, sob as mesmas condições, nem sempre é possível, mas em determinadas condições é possível determinar teoricamente o comportamento de algumas medidas feitas na amostra, como por exemplo a média. Mas isso depende, em grande parte, do procedimento (plano) adotado para selecionar a amostra. Assim, em problemas envolvendo amostras, antes de tomarmos uma decisão, teríamos de responder a quatro perguntas:

(a) Qual a população a ser amostrada?

(b) Como obter os dados (a amostra)?

(c) Que informações pertinentes (estatísticas) serão retiradas da amostra?

(d) Como se comporta(m) a(s) estatística(s) quando o mesmo procedimento de escolher a amostra é usado numa população conhecida?

Nas seções e capítulos subseqüentes tentaremos responder a essas perguntas.

## 10.4 Como Selecionar uma Amostra

As observações contidas em uma amostra são tanto mais informativas sobre a população quanto mais conhecimento explícito ou implícito tivermos dessa mesma população. Por exemplo, a análise da quantidade de glóbulos brancos obtida de algumas gotas de sangue da ponta do dedo de um paciente dará uma idéia geral da quantidade dos glóbulos brancos no corpo todo, pois sabe-se que a distribuição dos glóbulos brancos é homogênea, e de qualquer lugar que se tivesse retirado a amostra ela seria "representativa". Mas nem sempre a escolha de uma amostra adequada é imediata. Por exemplo, voltando ao Exemplo 10.2, para o qual queríamos obter uma amostra de habitantes para saber a opinião sobre um projeto governamental, escolhendo intencionalmente uma amostra de 200 indivíduos moradores de certa região beneficiada pelo projeto, saberemos de antemão que o resultado conterá um *viés de seleção*. Isto é, na amostra, a proporção de pessoas favoráveis ao projeto deverá ser maior do que no todo, donde a importância da adoção de procedimentos científicos que permitam fazer inferências adequadas sobre a população.

A maneira de se obter a amostra é tão importante, e existem tantos modos de fazê-lo, que esses procedimentos constituem especialidades dentro da Estatística, sendo *Amostragem* e *Planejamento de Experimentos* as duas mais conhecidas. Poderíamos dividir os procedimentos científicos de obtenção de dados amostrais em três grandes grupos:

(a) *Levantamentos Amostrais*, nos quais a amostra é obtida de uma população bem definida, por meio de processos bem protocolados e controlados pelo pesquisador. Podemos, ainda, subdividi-los em dois subgrupos: levantamentos probabilísticos e não-probabilísticos. O primeiro reúne todas aquelas técnicas que usam mecanismos aleatórios de seleção dos elementos de uma amostra, atribuindo a cada um deles uma probabilidade, conhecida *a priori*, de pertencer à amostra. No segundo grupo estão os demais procedimentos, tais como: amostras intencionais, nas quais os elementos são selecionados com o auxílio de especialistas, e amostras de voluntários, como ocorre em alguns testes sobre novos medicamentos e vacinas. Ambos os procedimentos têm suas vantagens e desvantagens. A grande vantagem das amostras probabilísticas é medir a precisão da amostra obtida, baseando-se no resultado contido na própria amostra. Tais medidas já são bem mais difíceis para os procedimentos do segundo grupo.

    Estão nessa situação os Exemplos 10.1 (conhecer os salários da Cia. MB), 10.2 (identificar a proporção de indivíduos favoráveis ao projeto), 10.4 (pesos dos pacotes de café) etc.

(b) *Planejamento de Experimentos*, cujo principal objetivo é o de analisar o efeito de uma variável sobre outra. Requer, portanto, interferências do pesquisador sobre o ambiente em estudo (população), bem como o controle de fatores externos, com o intuito de medir o efeito desejado. Podemos citar como exemplos aquele já citado sobre a altura de um produto na gôndola de um supermercado afetar as vendas e o Exemplo 10.6. Em ensaios clínicos em medicina, esse tipo de estudo é bastante usado, como por exemplo para testar se um novo medicamento é eficaz ou não para curar certa doença.

(c) *Levantamentos Observacionais*: aqui, os dados são coletados sem que o pesquisador tenha controle sobre as informações obtidas, exceto eventualmente sobre possíveis

erros grosseiros. As séries de dados temporais são exemplos típicos desses levantamentos. Por exemplo, queremos prever as vendas de uma empresa em função de vendas passadas. O pesquisador não pode selecionar dados, esses são as vendas efetivamente ocorridas. Nesses casos, a especificação de um modelo desempenha um papel crucial na ligação entre dados e população.

No caso de uma série temporal, o modelo subjacente é o de *processo estocástico*; podemos pensar que a série efetivamente observada é uma das *infinitas possíveis realizações desse processo*. A população hipotética aqui seria o conjunto de todas essas realizações, e a série observada seria a amostra. Veja Morettin e Toloi (2006) para mais informações.

Neste livro iremos nos concentrar principalmente em levantamentos amostrais e, mais ainda, num caso simples de amostragem probabilística, a *amostragem aleatória simples, com reposição*, a ser designada por AAS. O leitor poderá consultar Bussab e Bolfarine (2005) para obter mais detalhes sobre outros procedimentos amostrais. Um breve resumo sobre alguns planos é dado no Problema 37. Noções sobre planejamento de experimentos podem ser vistas em Peres e Saldiva (1982).

### Problemas

1. Dê sua opinião sobre os tipos de problemas que surgiriam nos seguintes planos amostrais:
   (a) Para investigar a proporção dos operários de uma fábrica favoráveis à mudança do início das atividades das 7h para as 7h30, decidiu-se entrevistar os 30 primeiros operários que chegassem à fábrica na quarta-feira.
   (b) Mesmo procedimento, só que o objetivo é estimar a altura média dos operários.
   (c) Para estimar a porcentagem média da receita municipal investida em lazer, enviaram-se questionários a todas as prefeituras, e a amostra foi formada pelas prefeituras que enviaram as respostas.
   (d) Para verificar o fato de oferecer brindes nas vendas de sabão em pó, tomaram-se quatro supermercados na zona sul e quatro na zona norte de uma cidade. Nas quatro lojas da zona sul, o produto era vendido com brinde, enquanto nas outras quatro era vendido sem brinde. No fim do mês, compararam-se as vendas da zona sul com as da zona norte.
2. Refazer o Problema 7 do Capítulo 8.

## 10.5 Amostragem Aleatória Simples

A amostragem aleatória simples é a maneira mais fácil para selecionarmos uma amostra probabilística de uma população. Além disso, o conhecimento adquirido com esse procedimento servirá de base para o aprendizado e desenvolvimento de outros procedimentos amostrais, planejamento de experimentos, estudos observacionais etc. Comecemos introduzindo o conceito de AAS de uma população finita, para a qual temos uma listagem de todas as N unidades elementares. Podemos obter uma amostra nessas condições, escrevendo cada elemento da população num cartão, misturando-os numa urna e sorteando tantos cartões quantos desejarmos na amostra. Esse procedimento torna-se inviável quando a população é muito grande. Nesse caso, usa-se um processo alternativo,