Capítulo 4

Estatística descritiva

Não é fácil descrever o que seja estatística, ou tudo o que ela compreende. Mas, podemos vê-la como um conjunto de técnicas que permitem colher dados, organizá-los e analisá-los, para que possamos deles extrair algum tipo de conhecimento que nos interesse.

Nesse capítulo vamos apresentar uma pequena fração dessas técnicas. A estatística descritiva, segundo dizem os especialistas, é a etapa inicial que tomamos para sumarizar e tentar entender os dados. Note que nosso objetivo aqui não é ensinar estatística ou como utilizá-la, mas sim mostrar como podem ser calculas algumas medidas que estão associadas à estatística descritiva.

Vamos supor, então, que temos uma população e que dela queremos estimar uma variável. Por exemplo, queremos saber qual a altura da população masculina no Brasil. Ou quantos litros de cerveja bebem as estudantes de engenharia da USP, por semana. Infelizmente, não podemos coletar esses dados de todos os elementos que compõem essas populações. Por isso, fazemos uma amostragem, por exemplo, entrevistando ou medindo um subconjunto pequeno desses elementos.

Assim, temos um conjunto de dados, x_1, x_2, x_3, ..., x_n, que contém os valores medidos para a variável de interesse, para n elementos que compõem a nossa amostra. Por exemplo, podemos entrevistar dez estudantes de engenharia e descobrir que, por semana, elas bebem a quantidade de cerveja mostrada na Tabela 4.1.

Tabela 4.1: Quantidade de cerveja ingerida pelas estudantes de engenharia da USP

Estudante	1	2	3	4	5	6	7	8	9	10
Litros	1,5	2	2	4	0	2,5	3,5	2	6	3,5

A primeira medida, talvez a mais conhecida, é a média desses dados. Ela é uma das medidas que chamamos de medidas de tendência central ou medidas de posição. Como o nome sugere, elas servem que tenhamos uma ideia da localização dos dados, dentro da escala em que foram medidos. A média pode ser calculada da seguinte maneira:

$$média = \frac{x_1 + x_2 + ... + x_n}{n}$$

Ou seja, o cálculo da média é feita somando-se todos os valores da amostra e dividindo a sua soma pelo número de elementos na amostra, n. No caso das meninas que bebem cerveja, se somarmos os valores das medidas de cada uma delas, obtemos 27 litros de cerveja por semana. Ou, a média de 2,7 litros.

Outra medida de posição é a mediana. Ela é definida como o valor central das medidas. Ou seja, metade dos valores da amostra é maior do que a mediana e metade é menor.

Para calcular a mediana fica mais fácil ordenar a nossa amostra. Se o tamanho da amostra for ímpar, a mediana corresponde ao valor central da sequência de valores ordenados. Se for par, fazemos a média dos dois elementos centrais. Por exemplo, ordenando os valores da nossa amostra exemplo, obtemos a sequência a seguir.

$$0 \quad 1{,}5 \quad 2 \quad 2 \quad 2 \;\big|\; 2{,}5 \quad 3{,}5 \quad 3{,}5 \quad 4 \quad 6$$

Como o tamanho da amostra é par, não existe um elemento central, então pegamos o quinto e o sexto elementos e calculamos a média deles. Assim, obtemos a mediana: $(2 + 2, 5)/2 = 2, 25$.

A moda corresponde ao valor que aparece mais vezes ou que é mais comum na amostra. Por exemplo, entre as estudantes de engenharia, o valor mais comum para consumo semanal de cerveja é de dois litros. Esse valor aparece três vezes, na amostra coletada. Para calcular a moda, precisamos saber qual é a "frequência" de cada valor da amostra, ou seja, quantas vezes cada valor apareceu. Como na Tabela 4.2, na qual vemos quantas vezes apareceu cada valor da amostra. O valor mais frequente é o 2 que, portanto, é a moda da mostra.

Tabela 4.2: Frequência de ocorrências de cada valor da Tabela 4.1

Valor	0	1,5	2	2,5	3,5	4	6
Frequência	1	1	3	1	2	1	1

Note que é possível que uma amostra tenha mais do que uma moda. Se, por exemplo, o valor 6 aparecesse, também, três vezes na amostra, teríamos dois valores para a moda: 2 e 6. Então, precisamos olhar na tabela de frequências, todos os valores da amostra que correspondem ao maior valor de frequência para escolher a moda.

Outras medidas, chamadas de dispersão, servem para verificar o quanto os dados variam. No nosso exemplo, se tivéssemos uma amostra em que todas as moças tomam exatamente 2,7 litros de cerveja por semana, continuaríamos com a mesma média mas com uma dispersão menor. Ou sejam, os dados variam menos do que na amostra original.

A primeira medida é a amplitude. Ela é calculada pela diferença entre o maior valor e o menor valor da amostra. Se a amostra estiver ordenada, basta subtrair o primeiro do último valor. No nosso exemplo, a amplitude é $6 - 0 = 6$.

Já o cálculo da variância é um pouco mais trabalhoso. A variância avalia a dispersão total medindo o quanto cada ponto se afasta da medida central da média. Em outras palavras, calculamos a variância como:

$$variância = \frac{(x_1 - média)^2 + (x_2 - média)^2 + ... + (x_n - média)^2}{n - 1}$$

Ou seja, para cada valor da amostra, computamos a sua diferença com a média e a elevamos ao quadrado. Somamos todos esses valores e dividimos a soma pelo tamanho da amostra menos um. O desvio padrão é uma outra medida de dispersão, definida apenas como a raiz quadrada da variância.