

Calculando mediana

Transcrição

Calculamos a média de duração dos cursos em dias e obtivemos um número em torno de 48, uma medida pontual e importante, mas que destoa um pouco do histograma que produzimos. Nele, a agregação de valores está entre 0 e 10 dias, e poderemos trabalhar melhor a média que, nesse caso, parece contradizer o gráfico.

Na amostra que a empresa forneceu, há dados de alunos que levaram mais de 500 dias para concluir um curso. A discrepância entre os valores distorce a média, que está sujeita a esse tipo de problema. Para resolver isso, utilizaremos outra medida estatística, a famosa **mediana** (`median`).

Para calculá-la, digitaremos no R Script:

```
median(duracao$dias)
```

A mediana representa o **valor central** da amostra, deixando metade dos dados ou dos casos para baixo ou à esquerda, e a outra metade à direita ou acima dela. Ao executarmos esse comando, teremos o mesmo problema encontrado ao tentarmos executar `mean` pela primeira vez:

```
> median(duracao$dias)
[1] NA
```

Ela não consegue trabalhar com dados indisponíveis (`NA`), e por isso precisaremos informar à mediana que eles devem ser removidos. Acrescentaremos, após vírgula (`,`), `na.rm = T` ao `median` :

```
median(duracao$dias, na.rm = T)
```

Teremos como retorno:

```
> median(duracao$dias, na.rm = T)
[1] 8
```

Notem que interessante: a mediana, nesse caso é 8. Olharemos o histograma para localizar, por exemplo, onde a mediana ficaria.

O número 8 estaria muito mais próximo de 0 do que de 20, no eixo `Tempo`. Por isso, intuitivamente, ela faz mais sentido. Se utilizássemos a média (48), ela ficaria mais próxima de 50 no eixo horizontal, cuja `Quantidade` foi reduzida significativamente, e a informação entraria em conflito com a que se obtém por meio do gráfico.

Passaríamos à empresa informações contraditórias, dizendo que a maioria dos valores estão entre 0 e 20 no eixo do `Tempo`, e que a média se aproxima a 50, o que não faria sentido. Por isso, uma mediana de 8 corresponde a um valor mais intuitivo, que combina com o histograma e corrobora suas informações, tornando a análise mais robusta.

