01

Calculando média

Transcrição

Há uma dispersão dos dados no histograma, que funciona bem para termos uma visão geral de como a amostra está distribuída. No entanto, não obtemos medidas pontuais por meio do gráfico. Por exemplo, se perguntarem a **média** de duração dos cursos em dias, não conseguiremos passar essa informação por meio do histograma.

Para obtermos essa medida estatística pontual, precisaremos solicitar a média de tempo que os alunos levam para completar um curso, diretamente ao programa. Para isso, utilizaremos a função estatística `mean`, "média" em inglês. Em seguida, e entre parênteses, digitaremos `duracao$dias` para especificar o banco de dados e a coluna em que estamos interessados, e assim, obtermos o número de dias que os alunos levam para completar o curso.

```
mean(duracao$dias)
```

Executaremos o código e, no Console, teremos como retorno:

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

`NA` é um problema, pois não representa média. É uma sigla para `Not Available`, "Não Disponível" em inglês. Ou seja, o programa não conseguiu calcular a média de dias, porque faltam dados na amostra - algumas linhas do banco de dados estão com a coluna de `dias` preenchida com a sigla `NA`.

Esse preenchimento pode ocorrer por dois motivos: o aluno desistiu no meio do caminho e não concluiu o curso, ou ele o concluiu depois de o cliente ter nos enviado a amostra. Como ainda não temos o número de dias que ele levaria para concluir, ao executarmos `mean`, temos `NA` como retorno. Precisaremos informar ao programa que as colunas preenchidas com `NA` devem ser ignoradas. Selecionaremos somente os casos em que os alunos já concluíram os cursos.

Para calcularmos a média ignorando os espaços preenchidos com `NA`, acrescentaremos no código `mean` o parâmetro `na.rm` após vírgula (`,`), que significa "remove not availables", "remover não disponíveis" em inglês. E igualaremos a verdadeiro (`TRUE` ou `T`). Assim, eliminaremos os dados indisponíveis no RStudio.

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

Ao executarmos esse comando do R Script, no Console teremos como retorno:

```
> mean(duracao$dias, na.rm = T)
[1] 47.83649
```

Obtivemos a média de duração em dias: `47.83649`. Em termos práticos, arredondando esse número, temos uma média de `48` dias para os alunos completarem um curso nessa empresa, a partir da amostra do banco de dados. A média é um dos primeiros e mais importantes conceitos estatísticos que utilizamos em diversas ocasiões.

Essa talvez seja a primeira informação que levaremos à empresa:

"Com relação à duração dos cursos, os seus alunos estão levando, em média, 48 dias para completá-los".