01

Calculando Média

Transcrição

Vimos uma dispersão dos dados no histograma.

Ele é bom 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 obter essa medida estatística pontual, precisamos solicitar a média de tempo que os alunos levam para completar um curso, diretamente ao programa.

Para isso, utilizamos 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, obter 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.

Isso acontece 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 concluiu o curso depois de o cliente ter nos enviado a amostra.

Como ainda não temos o número de dias que ele levaria para concluir, ao executar `mean`, temos `NA` como retorno.

Precisamos 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 calcular a média ignorando os espaços preenchidos com `NA`, acrescentaremos no código `mean`, após vírgula (`,`), o parâmetro `na.rm`, que significa "remove not availables", "remover não disponíveis" em inglês.

E igualaremos a verdadeiro (`TRUE` ou `T`). Assim, eliminamos os dados indisponíveis no RStudio.

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

Ao executar esse comando do R Script, no Console, o retorno será:

```
> 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".