05

Sumário Estatístico

Transcrição

Calculamos a média e a mediana da duração dos cursos em dias.

Existem diversas outras medidas estatísticas que poderíamos solicitar ao RStudio. Porém, é chato desenvolver uma por uma.

Existe uma forma mais eficiente de fazer isso. No programa, há um comando que fornece um **"Sumário Estatístico"** que calcula média e mediana, além de outras informações que ainda não vimos.

Por exemplo, o número máximo de dias que um aluno levou para concluir um curso. Tentaremos obter essa informação por meio do "Sumário Estatístico" (`summary`).

Para isso, no R Script, digitaremos em uma nova linha:

```
summary(duracao$dias)
```

Entre parênteses, especificamos o banco de dados (`duracao`) e a variável que estamos interessados (`dias`).

Ao executar esse comando, no Console, teremos o seguinte retorno:

```
> summary(duracao$dias)
    Min. 1st Qu.  Median   Mean 3rd Qu.    Max.    NA's
    0.00    2.00    8.00  47.84   45.00  538.00    3828
```

Foi produzida uma mini tabela. Na primeira linha dela, temos os títulos das medidas e na linha de baixo, as respectivas quantidades.

Algumas medidas são familiares, por exemplo, a média (`Mean`) com o valor (`47.84`) que calculamos individualmente. A mediana (`Median`) também é localizada com o mesmo valor (`8`).

Poderíamos ter executado esse sumário para obter a média e a mediana que calculamos anteriormente.

Mas, ele traz outros dados interessantes, como o número máximo de dias que um aluno levou para concluir um curso, apontado por `Max` , que na amostra enviada pela empresa foi de `538` dias, ou seja, mais de um ano.

Obtivemos outros parâmetros estatísticos, como o primeiro e o terceiro quartil. O **primeiro** refere-se ao valor que deixa **25%** dos casos **abaixo** dele e o **terceiro quartil** refere-se ao valor que deixa **25%** dos casos **acima** dele.

Na tabela, o valor do primeiro (`1st Qu.`) é `2.00` , indicando que em 25% dos casos os alunos levaram menos de `2` dias para concluir o curso.

O valor do terceiro (`3rd Qu.`) informa que 25% dos alunos levaram mais que `45` dias para concluir os cursos.

Por fim, em `Min.` temos o valor mínimo `0.00` que já sabíamos, pois não é possível um aluno levar menos que `0` dias para concluir um curso, e o valor de dados faltantes (`NA's`) que é `3828` , indicando que na amostra do banco de dados não temos informação sobre quantos dias, `3828` alunos levaram para concluir um curso.

Refraseando: `3828` alunos desistiram e não concluíram os cursos **ou** ainda não tinham concluído os cursos quando o banco de dados foi gerado.

Agora, se `3828` alunos não concluíram ou concluíram os cursos após o envio da amostra do banco de dados, qual a proporção disso em relação ao total de alunos?

Se isso for um problema, qual o tamanho dele? Para calcular, podemos usar funcionalidades do RStudio, que funciona como calculadora também, permitindo realizar diversas contas.

Para o cálculo dessa proporção, primeiro, precisamos descobrir o número de alunos do banco de dados por meio da dimensão (`dim`):

```
dim(duracao)[1]
```

Entre colchetes, inserimos `1` para orientar o programa a utilizar o banco de dados `duracao` e fornecer o tamanho da dimensão `1` , que é um número de linhas no banco de dados.

Ao executar esse comando, no Console, teremos:

```
> dim(duracao)[1]
[1] 6366
```

O número `6366` representa o tamanho da amostra, o número de matrículas. Podemos ter, por exemplo, um mesmo aluno contado mais de uma vez na mesma amostra.

Certamente, temos o mesmo curso repetidas vezes, também. Em `6366` temos a junção de número de matrículas. O dado individual é um aluno fazendo um determinado curso.

Precisamos dividir o número de dados indisponíveis pelo número total. Podemos calcular diretamente no R Script:

```
3828/6366
```

Para dividir valores, no RStudio, utilizamos barra (`/`). Ao executar a divisão, no Console, teremos a proporção que estamos buscando:

```
> 3828/6366
[1] 0.6013195
```

O resultado está em decimais (`0.6013195`). Ele indica que 60.13% das matrículas não concluíram os cursos por desistência, ou foram concluídas após o envio da amostra.

Como alguns cursos levam tempo até a conclusão, é natural depararmos com casos como esses.

Se necessário, levamos essa informação para a empresa, que saberá melhor como lidar com ela. O importante é informarmos que em 60.13% dos casos, não há conclusão.

Mas, ao que se refere essa porcentagem? Não sabemos as dimensões da amostra. Vamos colocar esse dado em comparação com o todo que estamos trabalhando.

Calcularemos a quantidade de cursos únicos para passar uma informação completa para a empresa, por meio de `length`. Como queremos calcular somente os casos únicos, utilizaremos `unique`:

```
length(unique(duracao$curso))
```

Ao executar o comando, no Console, teremos como retorno:

```
> length(unique(duracao$curso))
[1] 264
```

Ou seja, na amostrá há `264` cursos únicos.

Para finalizar, faremos a mesma coisa com os alunos, para definir a quantidade de alunos únicos:

```
length(unique(duracao$aluno))
```

O retorno, no Console, será:

```
> length(unique(duracao$aluno))
[1] 484
```

Poderíamos passar a informação que obtivemos para a empresa, da seguinte forma:

"Na amostra que vocês nos deram, há 6366 matrículas, 264 cursos únicos e 484 alunos únicos, dos quais 60.13% não concluíram os cursos por desistência ou concluíram somente após o envio da amostra do banco de dados."