05

Sumário estatístico

Transcrição

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

Como exemplo, usaremos o número máximo de dias que um aluno levou para concluir um curso. Tentaremos obter essa informação por meio do "Sumário Estatístico" (`summary`), digitando no R Script:

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

Entre parênteses, especificamos o banco de dados (`duracao`) e a variável em que estamos interessados (`dias`). Ao executarmos 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, e na sua primeira linha 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`) calculado individualmente. A mediana (`Median`) também é localizada com o mesmo valor (`8`). Poderíamos ter executado esse sumário para a obtenção da média e da mediana que calculamos anteriormente.

No entanto, 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. Já o valor do terceiro (`3rd Qu.`) indica que 25% dos alunos levaram mais de `45` dias para concluir os cursos. Por fim, em `Min.` temos o valor mínimo `0.00` , que já conhecíamos, pois não é possível um aluno levar menos que `0` dias para concluir um curso. E o valor de dados faltantes (`NA's`), `3828` , indica que na amostra do banco de dados não há informações sobre a quantidade de dias que `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 seu tamanho? Para calcularmos, podemos usar funcionalidades do RStudio, que também funciona como calculadora, realizando diversas contas.

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

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

Entre colchetes, inseriremos `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 executarmos esse comando, teremos no Console:

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

O número `6366` representa o tamanho da amostra, o número de matrículas. Poderemos ter, por exemplo, um mesmo aluno sendo contabilizado mais de uma vez na mesma amostra. Certamente temos o mesmo curso repetidas vezes, também. Em `6366`, há a junção de número de matrículas, e o dado individual indica um aluno fazendo um determinado curso. Precisaremos dividir o número de dados indisponíveis pelo número total. Podemos calcular isto diretamente no R Script:

```
3828/6366
```

Para dividir valores no RStudio, utilizamos barra (`/`).

Com isso, teremos a proporção que estamos buscando:

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

O resultado está em decimais (`0.6013195`), e significa que `60.13%` das pessoas matriculadas não concluíram os cursos por desistência, ou concluíram 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.

Mas a que se refere a porcentagem `60.13%`? Não sabemos as dimensões da amostra, portanto vamos colocar esse dado em comparação com o todo com que estamos trabalhando. Calcularemos a quantidade de cursos únicos para passarmos uma informação completa para a empresa, por meio de `length`. Como queremos calcular somente os casos únicos, utilizaremos `unique`:

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

Isto trará como retorno no Console:

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

Ou seja, na amostra há `264` cursos únicos. Por fim, faremos a mesma coisa para definirmos a quantidade de alunos únicos:

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

Cujo 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 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."