

Recapitulando os comandos

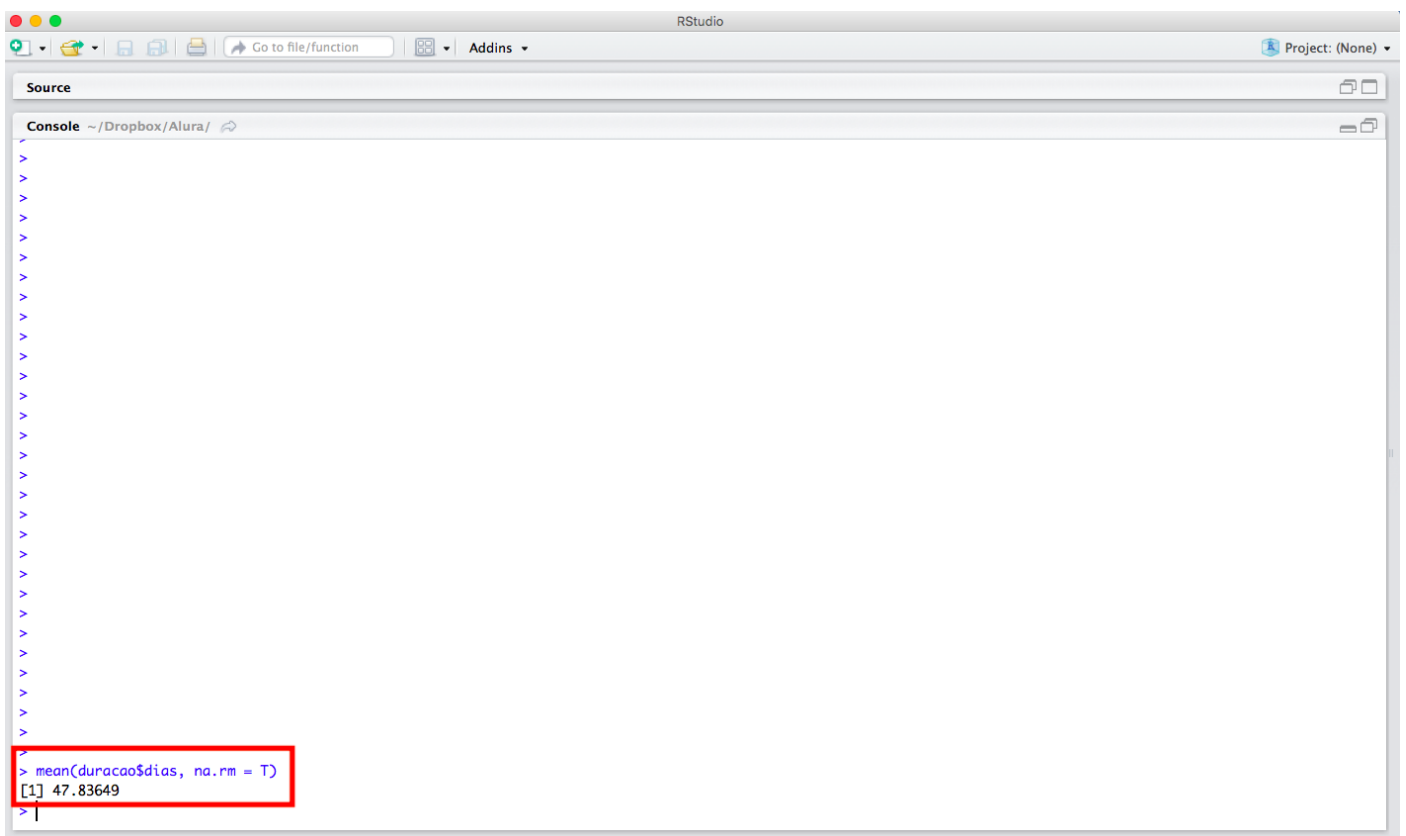
Nessa aula vimos como calcular algumas medidas de estatística descritiva da duração em dias dos cursos. Numa análise, isso é muito importante para termos a real noção do conjunto dos dados. Vamos recalculer essas estatísticas e analisá-las?

- Calcule a média de duração em dias dos cursos;
- Calcule a mediana da duração em dias dos cursos;
- Exiba um sumário estatístico de todo o banco *duracao*.

Opinião do Instrutor

- Calculando a média de duração em dias dos cursos:

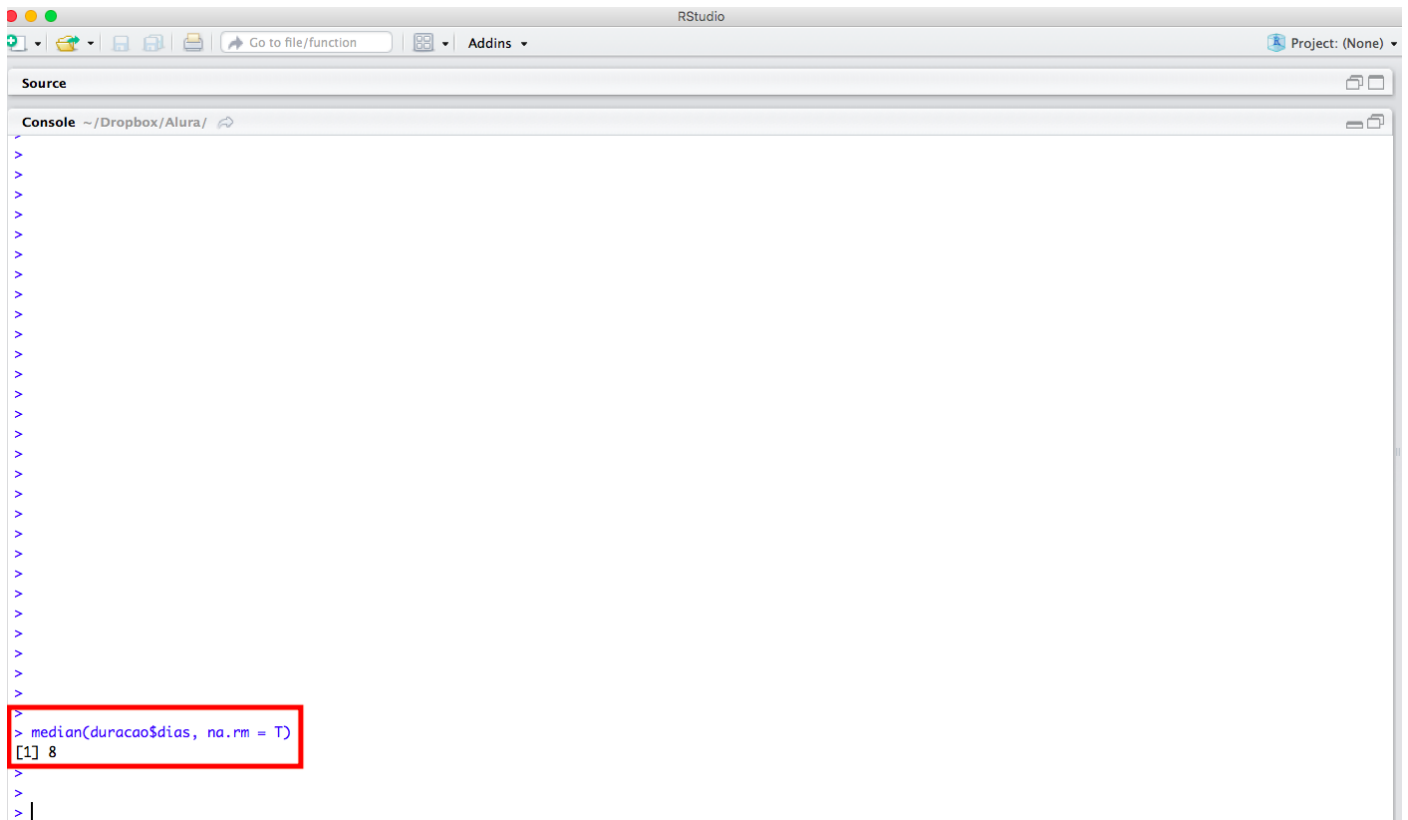
```
mean(duracao$dias, na.rm = T)
```

A screenshot of the RStudio interface. The top bar shows 'RStudio' and 'Project: (None)'. The 'Source' pane is empty. The 'Console' pane shows a series of prompt characters '>' and the command '> mean(duracao\$dias, na.rm = T)' which has been executed, resulting in the output '[1] 47.83649'. The command and its output are highlighted with a red rectangular box.

Veja que tivemos que especificar `na.rm = T` por causa dos dados faltantes. A média é a primeira e talvez mais importante estatística pontual. No entanto, ela pode nos dar uma visão errônea dos dados. Note que a média que encontramos foi aproximadamente 47,84 dias. Mas será que a maioria dos alunos leva tudo isso para concluir um curso?

- Calculando a mediana da duração em dias dos cursos:

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

The image shows a screenshot of the RStudio interface. The top bar includes the RStudio logo, a 'Go to file/function' search bar, and a 'Project: (None)' dropdown. The main window is divided into two panes: 'Source' and 'Console'. The 'Console' pane shows the command prompt with several lines of '>' indicating previous commands. The current command being executed is `median(duracao$dias, na.rm = T)`, which is highlighted with a red rectangular box. The output of this command is `[1] 8`, also within the red box. The console path is shown as `~/Dropbox/Alura/`.

Olha que resultado bastante diferente! A mediana `8` indica que metade dos alunos levam menos de 8 dias para concluir seus cursos e a outra metade leva mais que isso; ou seja, um valor bem distante daqueles 48 dias de média. A mediana parece nos dar uma visão mais apropriada nesse caso, quando comparamos com o histogramas que vimos anteriormente. A grande maioria dos dados está no começo do gráfico. Voltando à média, ela tem esse valor bem mais alto porque há casos discrepantes que *distorcem* a média, puxam a média para cima.

- Exibindo um sumário estatístico de todo o banco `duracao`:

```
summary(duracao)
```

Veja que com esse comando exibimos todas as estatísticas para todas as colunas. Mas, na verdade, as duas primeiras colunas dessa exibição não fazem sentido, pois os dados das variáveis `aluno` e `curso` são apenas os códigos dessas observações no banco de dados. Mas a coluna `dias` sim traz as informações úteis para nós: quartis, mediana, média e número de casos faltantes.

RStudio

Go to file/function

Addins Project: (None)

Source

Console ~/

```
>
>
>
> summary(duracao)
```

aluno	curso	dias
Min. : 1.0	Min. : 2.0	Min. : 0.00
1st Qu.:135.0	1st Qu.: 58.0	1st Qu.: 2.00
Median :291.0	Median : 96.0	Median : 8.00
Mean :271.3	Mean :112.9	Mean : 47.84
3rd Qu.:402.0	3rd Qu.:170.0	3rd Qu.: 45.00
Max. :500.0	Max. :276.0	Max. :538.00
		NA's :3828

```
>
```