

Medidas de Tendência Central



COMIXPLAIN

Essa história em quadrinhos foi criada no projeto de pesquisa Comixplain, financiado pela Innovation Call 2022 da Universidade de Ciências Aplicadas St. Pölten, Áustria.

Equipe:

Victor-Adriel De-Jesus-Oliveira
Hsiang-Yun Wu
Christina Stoiber
Magdalena Boucher
Alena Ertl

Contato:

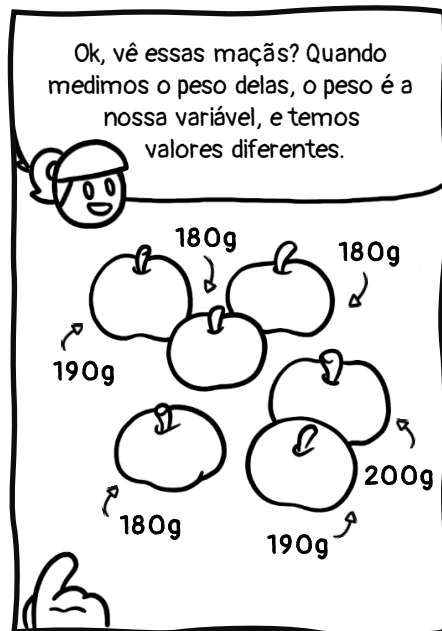
victor.oliveira@fhstp.ac.at

Ilustrações:

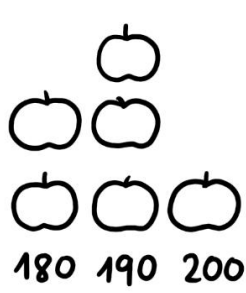
Magdalena Boucher & Alena Ertl



<https://fhstp.github.io/comixplain>

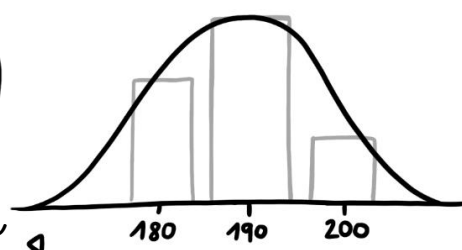


A melhor maneira de descrever uma variável é relatar os valores e a frequência em que eles aparecem. Isso é chamado de **DISTRIBUIÇÃO** da variável.



180 190 200

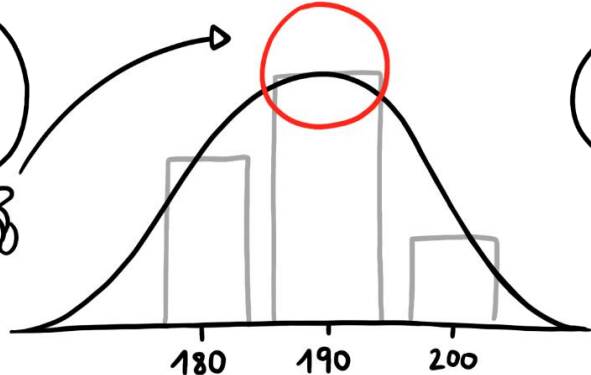
Se visualizarmos o peso das nossas maçãs, a distribuição seria assim, já que todas as maçãs desta cesta pesam aproximadamente o mesmo.



Peso das maçãs

Tem a forma de um sino...

Exato! Se houvesse outras maçãs na cesta, nosso melhor palpite para seu peso seria um valor em torno daquele ponto.



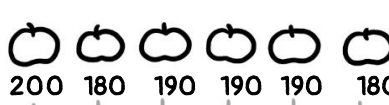
Ah, entendi. Mas... como eu calcularia esse valor?

Este ponto da nossa distribuição representa bem nossos dados – nós o chamamos **TENDÊNCIA CENTRAL**.

Podemos resumir a tendência central de várias maneiras. A **MÉDIA** é a maneira mais comum, e também é fácil de calcular.




Estas são as nossas seis maçãs:



200 180 190 190 190 180

Para calcular a média, somamos todos os valores de peso...



$200 + 180 + 190 + 190 + 190 + 180$

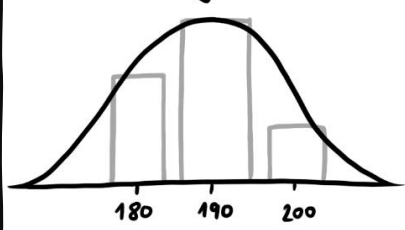
... e depois dividir pelo número de maçãs que temos...

$(200 + 180 + 190 + 190 + 190 + 180)$

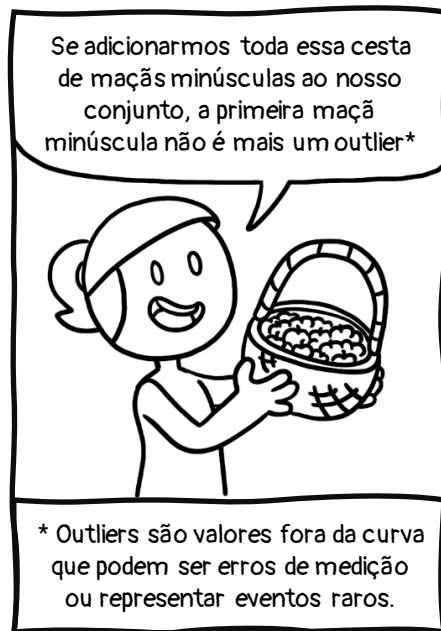
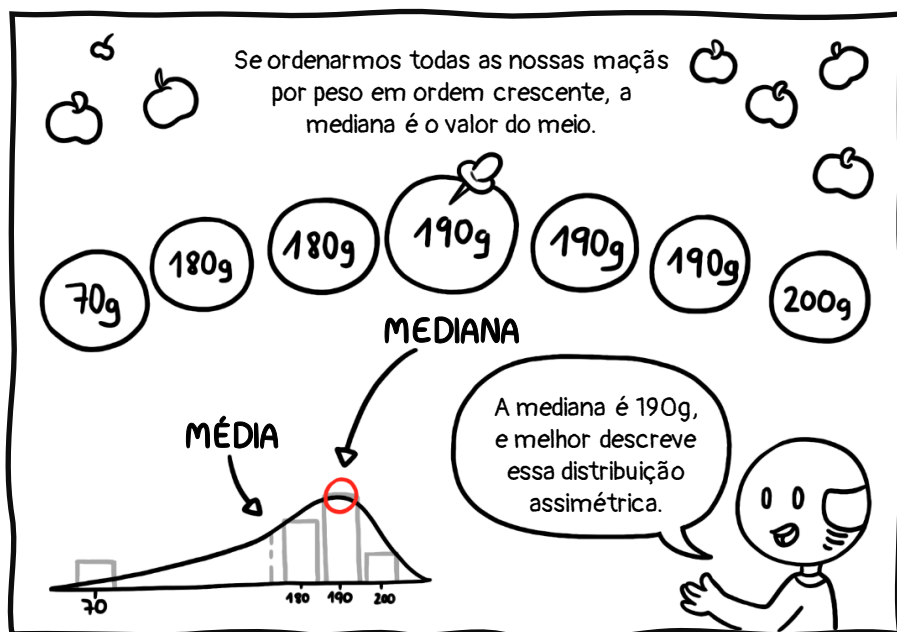
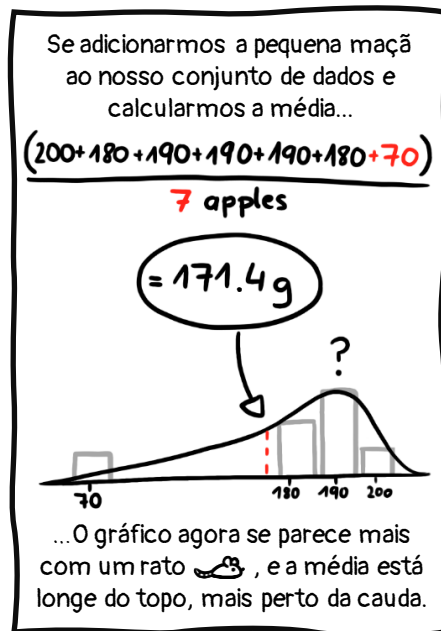
6

Isso nos dá um valor médio de

188.3g



Ah, isso é quase o topo do nosso gráfico em forma de sino!



Neste caso, nosso gráfico de repente tem dois picos:

Como ratos apaixonados!

Ah... Tá.

De qualquer forma, podemos usar algo chamado **MODA** para descrever a tendência central da nossa distribuição se esta tiver vários picos.

Peso das maçãs grandes e pequenas

A moda define o(s) valor(es) que ocorre(m) com mais frequência em um conjunto de dados.

Neste caso, temos várias modas. Mas algumas vezes existirá apenas uma ou mesmo nenhuma moda.

Você pode aplicar média, mediana e moda a diferentes amostras de maçãs. Mas, muitas vezes, algumas medidas representarão os dados melhor do que outras.

180, 180, 190, 190, 190, 200

70, 180, 181, 190, 191, 191, 200

60, 70, 70, 70, 80, 180, 180, 190, 190, 190, 200

M = 188.3 **MD = 190** **Mode = 190** **Melhores parâmetros**

M = 171.8 **MD = 190** **Mode = 191** **Melhores parâmetros**

M = 134.5 **MD = 180** **Mode = 70 & 190** **Melhor parâmetro**

Ok, obrigado... Aprendi muito! Agora só tenho que aplicar aos dados que tenho para apresentar. São de um aplicativo que rastreia medições de frequência cardíaca.

User ID	Heart Rate (bpm)	Time of Use	User Rating
1	45	13:00	1
2	50	9:00	5
3	55	10:00	3
4	57	9:00	4
5	63	14:00	5
6	70	15:00	5
7	65	16:00	4
8	75	15:00	2

Isso é factível. Olhe para seus dados e siga os exemplos das maçãs! Você pode usar a próxima página para seus cálculos.



Antes de virar a página, tente calcular média, mediana e moda para cada variável, e verifique qual parâmetro é mais adequado para descrever a tendência central.

Você pode fazer anotações nesta página!

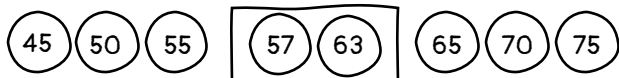
Fique à vontade para conferir teus cálculos.
Você pode fazer mais anotações nesta página!

FREQUÊNCIA CARDÍACA

Calculando a MÉDIA:

$$\frac{45+50+55+57+63+70+65+75}{8 \text{ usuários}} = \frac{480}{8} = 60 \text{ bpm}$$

Calculando a MEDIANA:



Se houver dois valores centrais, a média dos dois valores é a mediana:
 $(57+63)/2 = 60 \text{ bpm}$

Calculando a MODA:

45, 50, 55, 57, 63, 70, 65, 75

Cada valor só existe uma vez –
 a moda não existe!

Se a distribuição dos valores for simétrica, sem distorções, a média é geralmente igual à mediana.

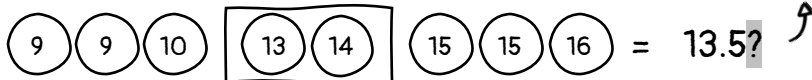


TEMPO DE USO MAIS FREQUENTE

Calculando a MÉDIA:

$$\frac{9+9+10+13+14+15+15+16}{8 \text{ usuários}} = \frac{101}{8} = 12.6?$$

Calculando a MEDIANA:



O tempo de uso não é um valor quantitativo – portanto, calcular média e mediana não faz nenhum sentido!

Calculando a MODA:

9:00, 10:00, 13:00, 14:00, 15:00, 16:00
 2x 1x 1x 1x 2x 1x = 2 modas:
 = 9:00 & 15:00

Moda não é adequada apenas para distribuições multimodais, mas também ao trabalhar com dados ordinais e categóricos.

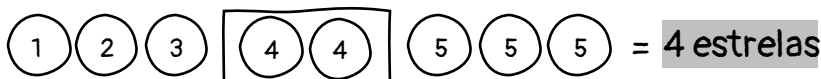


CLASSIFICAÇÃO POR ESTRELAS

Calculando a MÉDIA:

$$\frac{1+2+3+4+4+5+5+5}{8 \text{ usuários}} = \frac{29}{8} = 3.6 \text{ estrelas}$$

Calculando a MEDIANA:



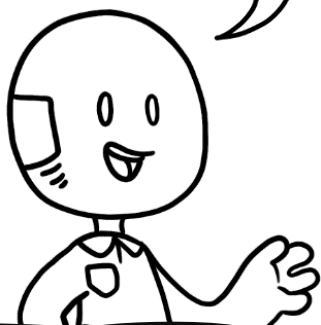
Calculando a MODA:

1 2 3 4 5
 1x 1x 1x 2x 3x = 5 estrelas

Para conjuntos de dados com distribuição assimétrica, a mediana é uma maneira melhor de descrever a tendência central.



Linguagens de programação, como **R**, te ajudam a calcular a tendência central de atributos em grandes conjuntos de dados. Com bibliotecas **R**, como o **tidyverse**, você pode visualizar rapidamente a distribuição de dados.



	model	year	hwy
1	jetta	1999	44
2	corolla	2008	37
3	civic	2008	36
4	civic	2008	36
5	corolla	1999	35
6	altima	2008	32
7	sonata	2008	31
+ outros 227 itens			

No tidyverse, você tem acesso a conjuntos de dados, como **mpg** com dados de economia de combustível. Ele inclui 11 atributos, como modelo do carro (**model**), ano de fabricação (**year**) e milhas rodoviárias por galão (**hwy**).

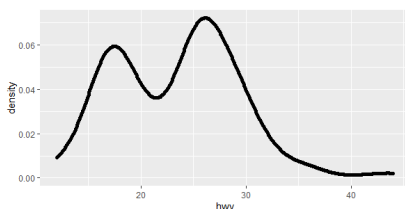


Você pode usar o **ggplot**, que está incluído no tidyverse, para visualizar a distribuição de dados de milhas rodoviárias por galão (**hwy**) usando um histograma, uma curva de densidade ou ambos.

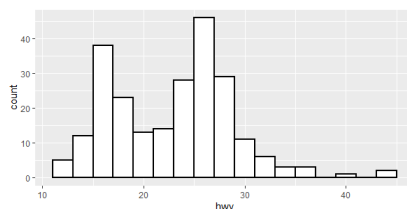


`install.packages("tidyverse")` # Instale-o apenas na primeira vez que usar a biblioteca

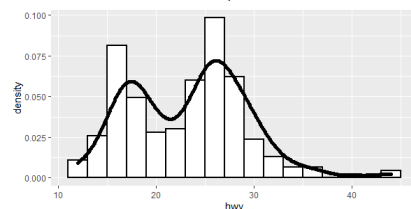
```
library(tidyverse)
plot <- ggplot(mpg, aes(x=hwy))
plot +
  geom_density()
```



```
plot +
  geom_histogram(
    colour="black",
    fill="white" )
```



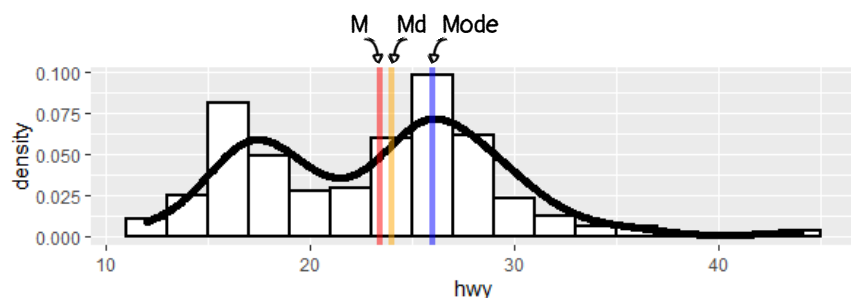
```
plot +
  geom_histogram(aes(y=..density..),
    colour="black",
    fill="white" ) +
  geom_density()
```



`mean(mpg$hwy)` = 23.4

`median(mpg$hwy)` = 24

`library(modeest)`
`mlv(mpg$hwy)` = 26



R inclui funções nativas para calcular média e mediana. Para a moda, você pode criar sua própria função ou usar a Most Likely Values (**mlv**) da biblioteca **modeest**.



Fontes:

Downey, A. (2014). Think stats: exploratory data analysis. O'Reilly Media, Inc.

Field, A. (2022). An adventure in statistics: The reality enigma. Sage.