

Estatística e Informática

Aula 05 - Medidas Estatísticas

Alan Rodrigo Panosso alan.panosso@unesp.br

Departamento de Engenharia e Ciências Exatas FCAV/UNESP

(13-05-2021)

Medidas de Posição ou de Tendência Central

Medidas de posição

São respostas breves e rápidas que sintetizam a informação não de uma forma de completa descrição dos dados ou eventual modelagem.

Essas medidas, portanto, fornecem a posição na reta real, informa sobre a posição de um conjunto de dados. As Principais medidas são:

1. Média:

- **Aritmética**
- Ponderada
- Geométrica (Apostila Apenas)
- Harmônica (Apostila Apenas)

2. Mediana

3. Moda

Média Populacional

Para a população, a média é definida como:

$$\mu = \frac{\sum_{i=1}^N x_i}{N}, \text{ com } i = 1, 2, \dots, N.$$

Onde N é o tamanho da população (e geralmente não a conhecemos).

Média amostral

É a mais utilizada das medidas de posição. A média aritmética de um conjunto de n observações da variável aleatória X , é o quociente da divisão por n da soma dos valores das observações dessa variável. A média amostral \bar{x} é a estimativa mais **eficiente**, **imparcial** e **consistente** da média da população μ .

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \text{ com } i = 1, 2, \dots, n.$$

Onde n é o tamanho da amostra.

O exemplo abaixo demonstra o cálculo da média da amostra da idade em anos de 49 alunos (idade_anos). Observe que a média tem as mesmas unidades de medida que as observações individuais.

Exemplo: Para os dados de idade_anos, da base de **Dados** da turma os valores de média.

Carregando Pacotes e a Base de Dados

```
library(readxl)
library(tidyverse)
dados_turmas <- read_excel("C:/Users/Usuario/Downloads/dados_turmas.xlsx")
View(dados_turmas)
```

Cálculo da Média Aritmética

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{19 + 20 + \dots + 20}{49} = 20,9184 \text{ anos}$$

```
x<-dados_turmas$idade_anos  
mean(x)
```

```
#> [1] 20.91837
```

Cálculo da Média Aritmética pela Frequência Relativa

$$\bar{x} = \sum_{i=1}^k f_i x_i$$

$$\bar{x} = 18(0,16) + 19(0,28) + \dots + 49(0,02)$$

$$\bar{x} = 20,9184 \text{ anos}$$

Tabela de Frequência para Idade

idade_anos	ni	fi
18	8	0.1632653
19	14	0.2857143
20	11	0.2244898
21	3	0.0612245
22	3	0.0612245
23	6	0.1224490
24	1	0.0204082
27	1	0.0204082
28	1	0.0204082
49	1	0.0204082

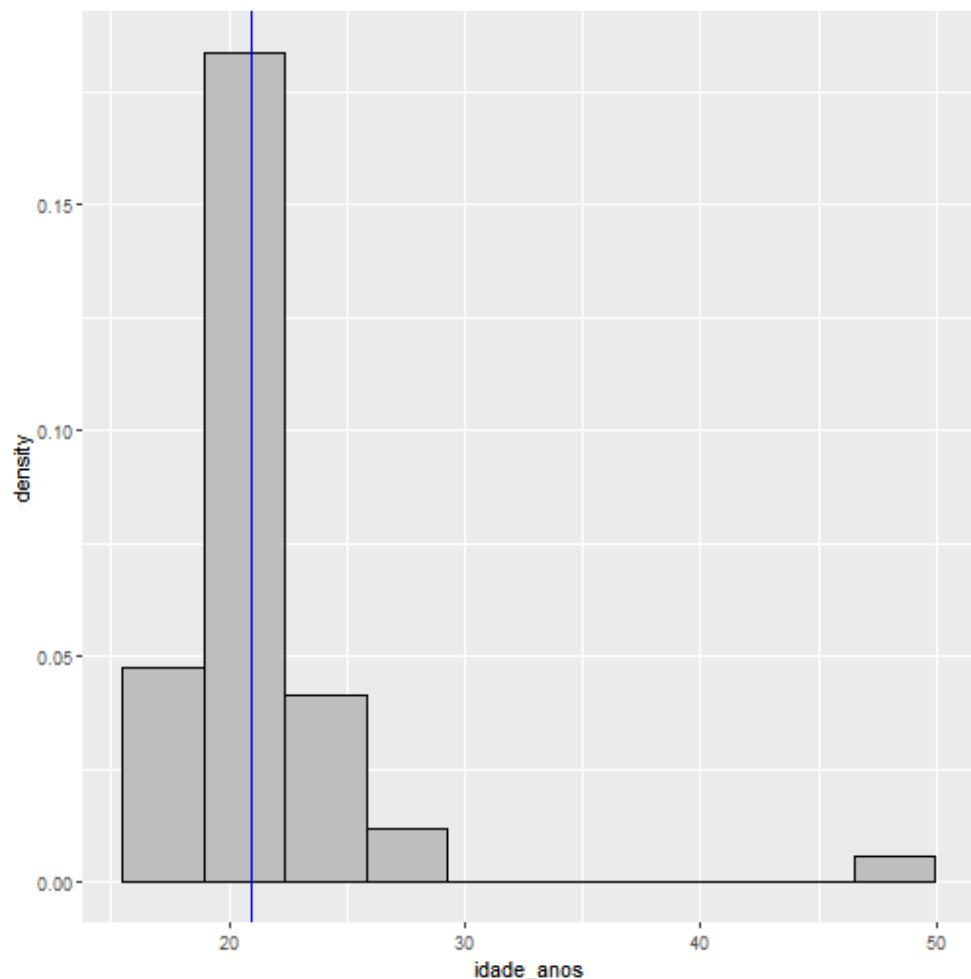
Hitograma e média para idade_anos

```
dados_turmas %>%  
  ggplot(aes(x=idade_anos, y=..density..)) +  
  geom_histogram(bins=10,  
                 color="black",  
                 fill="gray")+  
  geom_vline(xintercept = 20.9184,  
             color="blue")
```

Se os dados são plotados como um histograma a média é o centro de gravidade do histograma (linha azul).

Ou seja, se o histograma fosse feito de um material sólido, ele se equilibraria horizontalmente com o ponto de apoio em \bar{x} .

A média é aplicável aos dados de quantitativos, e **não deve ser usado para dados ordinais e nominais**.



Hitograma e média para altura

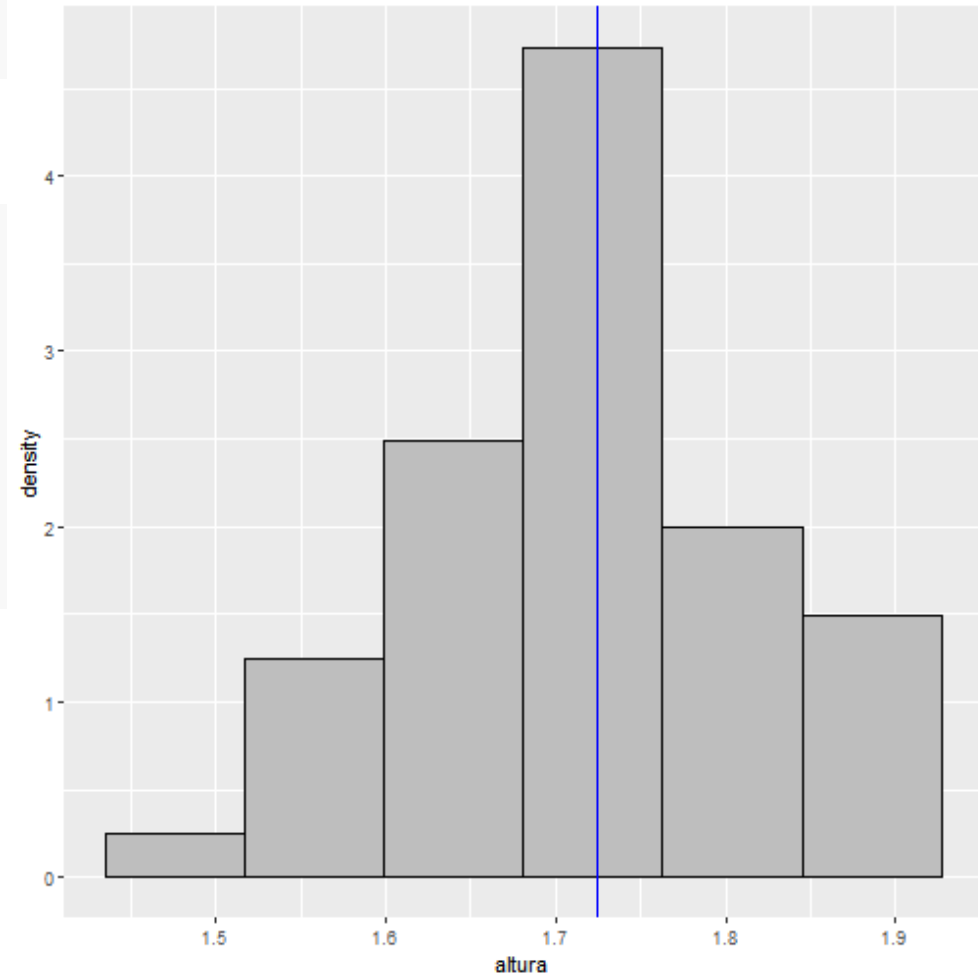
```
mean(dados_turmas$altura)
```

```
#> [1] 1.724286
```

```
dados_turmas %>%  
  ggplot(aes(x=altura, y=..density..)) +  
  geom_histogram(bins=6,  
                 color="black",  
                 fill="gray")+  
  geom_vline(  
    xintercept = 1.7242,  
    color="blue")
```

$$\bar{x} = \frac{1,98+1,59+\dots+1,78}{49}$$

$$\bar{x} = 1,724 \text{ m}$$



Observe o histograma e a média para a variável `altura`, e compare sua simetria com o histograma da variável `idade_anos`.

Mediana

É o valor que ocupa a posição central do conjunto de dados ordenados (X' ou X_o). Portanto, inicialmente as observações são ordenadas de acordo com a sua ordem de magnitude.

o valor da mediana é precedido e seguido pelo mesmo número de observações. E a mediana amostral é a melhor estimativa da mediana populacional quando o número de observações é grande.

A mediana será encontrada ou calculada em função do número de observações n presentes na base de dados, se acaso n for par ou ímpar:

$$\text{Se } n \text{ é ímpar: } Md = X_o\left(\frac{n+1}{2}\right)$$

$$\text{Se } n \text{ é par: } Md = \frac{X_o\left(\frac{n}{2}\right) + X_o\left(\frac{n}{2} + 1\right)}{2}$$

Exemplo 1:

Dado $X = \{50, 60, 20, 50, 30, 90, 70\}$, teremos a sua versão ordenada X_o dada por:

$X_o = \{20, 30, 50, 50, 60, 70, 90\}$, então,

$n = 7$, ímpar.

$$Md = X_o\left(\frac{n+1}{2}\right) = X_o\left(\frac{7+1}{2}\right) = X_o_4 = 50$$

No R:

```
X=c(20, 30, 50, 50, 60, 70 ,90)  
median(X)
```

```
#> [1] 50
```

Observe que a unidade da mediana é a mesma dos dados individuais.

No Excel:

DIAS360				fx				=MED(B2:B8)			
	A	B	C								
1		X									
2		50									
3		60									
4		20									
5		50									
6		30									
7		90									
8		70									
9	Mediana	50	=MED(B2:B8)								

Media e Mediana

Em uma distribuição simétrica (como a da variável `altura`), a mediana amostral também é uma estimativa imparcial e consistente de média populacional μ , contudo não é tão eficiente como a média amostral \bar{x} .

```
median(dados_turmas$altura)
```

```
#> [1] 1.74
```

```
dados_turmas %>%  
  ggplot(aes(x=altura, y=..density..)) +  
  geom_histogram(bins=6,  
                 color="black",  
                 fill="gray")+  
  geom_vline(  
    xintercept = c(1.7242,1.74),  
    color=c("blue","red"))
```

$Md = 1,74 \text{ m}$

Se a distribuição de frequência é assimétrica (como a da variável `idade_anos`) a mediana é uma pobre estimativa da média populacional.

Exemplo 2: Calcular a média e a mediana dos dados de salário (nº salário mínimos), que diz respeito à uma amostra de 6 colaboradores de uma empresa.

$$S = \{5, 2, 3, 6, 10, 9\}$$

$n = 6$, par.

$$Md = \frac{Xo_{(\frac{n}{2})} + Xo_{(\frac{n}{2}+1)}}{2}$$

Assim, para os dados de salários temos:

Vetor ordenado: $So = \{2, 3, 5, 6, 9, 10\}$, então a mediana será dada por:

$$Md = \frac{Xo_{(\frac{n}{2})} + Xo_{(\frac{n}{2}+1)}}{2} = \frac{Xo_{(3)} + Xo_{(4)}}{2} = \frac{5+6}{2} = 5,5 \text{ salários mínimos.}$$

A média será dada por:

$$\bar{x} = \frac{2+3+5+6+9+10}{6} = 5,83 \text{ salários mínimos}$$

Imagine que, ao invés de 10, o verdadeiro valor de ganho do 5º colaborador amostra seja 100, calcule a média e a mediana novamente para essa variável:

$$S = \{5, 2, 3, 6, 100, 9\}$$

$n = 6$, par.

$$Md = \frac{X_{o(\frac{n}{2})} + X_{o(\frac{n}{2}+1)}}{2}$$

Assim, para os dados de salários temos:

Vetor ordenado: $So = \{2, 3, 5, 6, 9, 100\}$, então a mediana será dada por:

$$Md = \frac{X_{o(\frac{n}{2})} + X_{o(\frac{n}{2}+1)}}{2} = \frac{X_{o(3)} + X_{o(4)}}{2} = \frac{5+6}{2} = 5,5 \text{ salários mínimos.}$$

A média será dada por:

$$\bar{x} = \frac{2+3+5+6+9+100}{6} = 20,83 \text{ salários mínimos}$$

Concluimos então que a mediana não é afetada por observações muito grandes ou muito pequenas, enquanto que a presença de tais extremos tem um significativo efeito sobre a média.

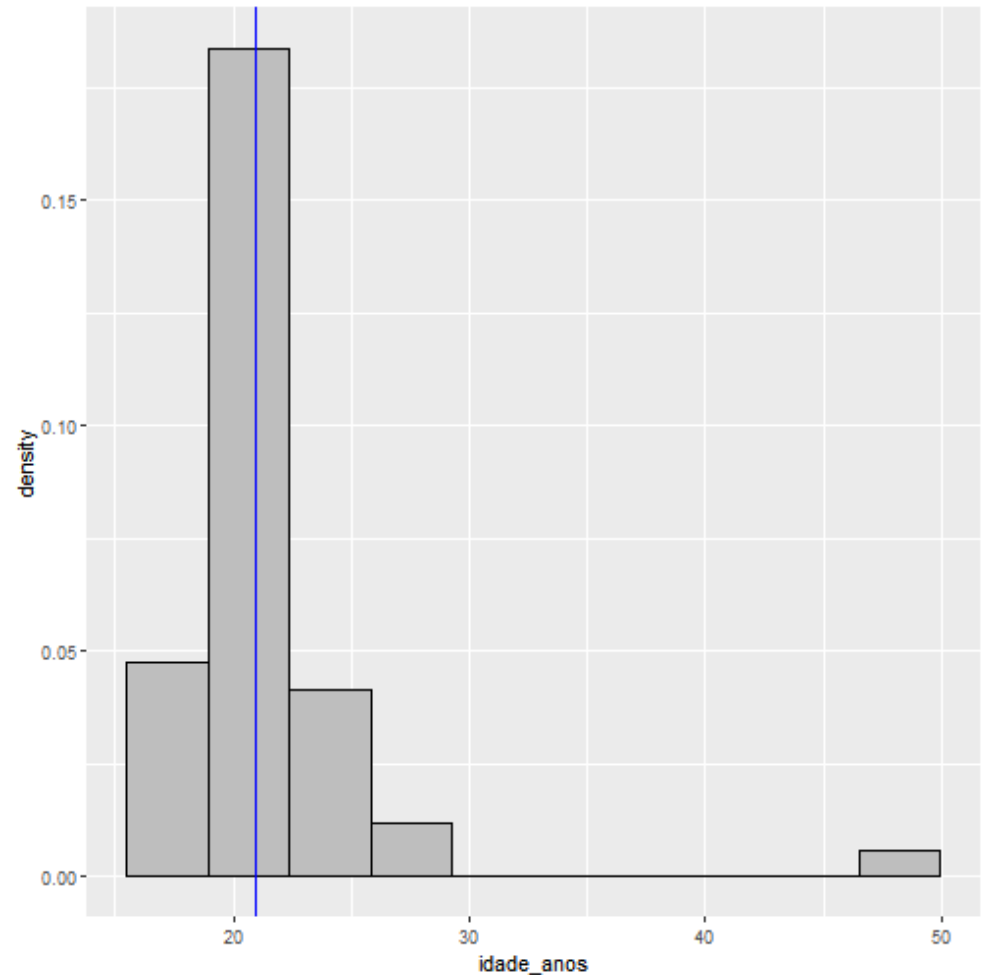
Moda

A moda é comumente definida como a medição que ocorre com mais frequência em um conjunto de dados.

Para os dados de `idade_anos` o valor mais presente do conjunto de dados é 19 anos.

Portanto, podemos ter distribuições com mais de um valor mais frequente (plurimodal), ou mesmo sem moda (amodal).

Outra forma de definir a moda como uma medida de concentração relativamente grande, pois algumas distribuições de frequência podem ter mais de um ponto de concentração, embora essas concentrações possam não conter precisamente as mesmas frequências.



```
x<-c(6, 7, 7, 8, 8, 8, 8, 8, 8, 9, 9, 10, 11, 12, 12, 12, 12, 12, 13, 13, 14)
tibble(x) %>%
  ggplot(aes(x=x,y=..density..)) +
  geom_histogram(bins = 9,
                 color="black",
                 fill="lightgray")
```

Quantis

É a extensão da noção de mediana, ou seja, são observações que dividem o conjunto ordenado de dados.

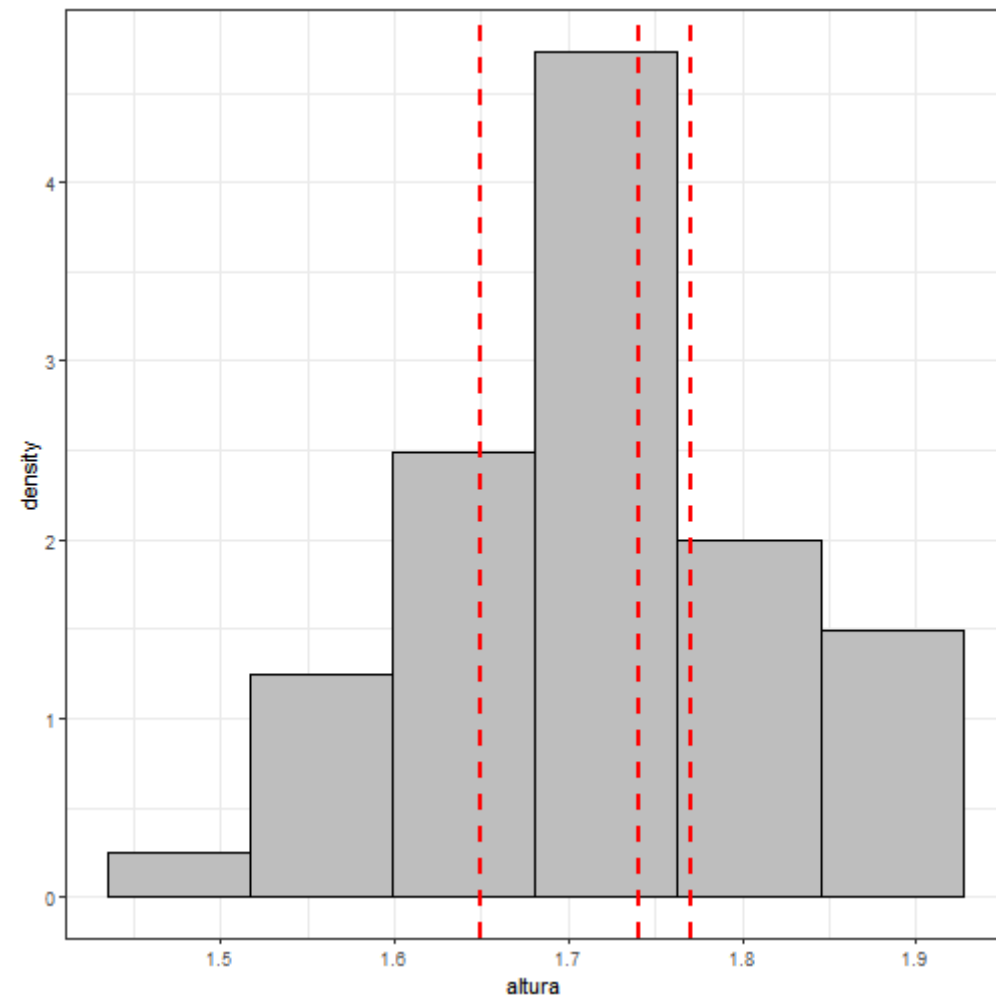
- Os Quantis de ordem 25, 50, 75 são chamados **Quartis** (Q1, Q2, Q3). Naturalmente, Q2 = mediana (Md).
- Os **Decis** são os quantis de ordem 10, 20, ..., 90 (D1, D2, ..., D9)
- Os **Percentis** são os quantis de ordem 1, 2, ..., 99 (P1, P2, ..., P99).

No R, eles podem ser encontrados por meio da função `quantile()` que tem como argumento o vetor de dados e a proporção dos dados acumulada até o respectivo valor.

```
altura<-dados_turmas$altura
q1 <- quantile(altura,0.25)
q2 <- quantile(altura,0.50)
q3 <- quantile(altura,0.75)
```

A função `geom_vline()`, ao lado, cria as linhas verticais na posição dos respectivos quartis no gráfico a seguir.

```
dados_turmas %>%
  ggplot(aes(x=altura,y=..density..)) +
  geom_histogram(bins=6,
                 col="black",
                 fill="gray") +
  geom_vline(xintercept=c(q1,q2,q3),
             color="red",lwd=1,
             linetype=2)+
  theme_bw()
```

Medidas de Dispersão ou de Variabilidade

Medidas de Dispersão:

O resumo de um conjunto de dados, por meio de uma única medida representativa de posição central, esconde toda informação sobre a variabilidade do conjunto de valores.

As medidas de variação medem o grau com que os dados tendem a se distribuir em torno de um valor central.

- Amplitude total
- Variância
- Desvio Padrão
- Erro Padrão da Média
- Coeficiente de Variação

Coeficientes de formas de distribuição:

- Coeficiente de Assimetria
- Coeficiente de Curtose

Exemplo:

Dados o conjunto de 4 amostras da mesma variável (altura de planta em cm), calcule a média e a Amplitude amplitude total de cada amostra.

X1	X2	X3	X4
9	7	0.6	0.6
9	8	3.4	9.0
9	9	9.8	9.0
9	10	13.8	9.0
9	11	17.4	17.4

Tabela de Médias $\left(\sum_{i=1}^n x_i\right)/n$

media_X1	media_X2	media_X3	media_X4
9	9	9	9

Tabela de Amplitudes $\left(\Delta = \text{Máximo} - \text{Mínimo}\right)$

Delta_X1	Delta_X2	Delta_X3	Delta_X4
0	4	16.8	16.8

Tabela Médias

media_X1	media_X2	media_X3	media_X4
9	9	9	9

Tabela Amplitudes

Delta_X1	Delta_X2	Delta_X3	Delta_X4
0	4	16.8	16.8

- Apesar das amostras apresentarem o mesmo valor médio, a terceira (X3) e quarta (X4) amostras são mais dispersas.
- As amostras X3 e X4, apesar do mesmo valor de média e amplitude, são bem distintas. Isso corre pois a Amplitude leva em consideração apenas **dois** valores do conjunto de dados, o conveniente seria considerar uma medida que utiliza-se todas as observações.

Desvios ou erros (e_i)

Para resolvermos o problema da amplitude consideramos os desvios ou erros (e_i)'s de cada observação em relação a um ponto de referência, no caso, a média aritmética do conjunto de dados, portanto, os desvios de cada observação é dado por:

$$e_i = x_i - \bar{x}$$

Para os dados da amostra 3 (X3) temos:

X3	Desvios
0.6	-8.4
3.4	-5.6
9.8	0.8
13.8	4.8
17.4	8.4

Como demonstrado anteriormente, a soma dos erros e_i é sempre igual a zero:

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x}$$

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n \cdot \bar{x}$$

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n \cdot \frac{\sum_{i=1}^n x_i}{n}$$

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n x_i = 0$$

Soma dos Quadrados dos Desvios (SDQ)

Para evitar a inconveniência do somatório dos desvios ser igual a zero, para qualquer conjunto de dados, elevamos ao quadrado cada um dos valores de desvios e_i .

$$e_i^2 = (x_i - \bar{x})^2$$

e temos

$$SQD = \sum_{i=1}^n (x_i - \bar{x})^2$$

A SQD leva a unidade dos dados ao quadrado e pode ser calculada no R:

```
X3<-c(0.6,3.4,9.8,13.8,17.4)
QD <- (X3-mean(X3))^2
QD
```

```
#> [1] 70.56 31.36 0.64 23.04 70.56
```

```
sum(QD)
```

```
#> [1] 196.16
```

No Excel temos:

	A	B	C	D	E	F
		X3	ei	ei ²		
2		0.6	=B2-\$B\$7	70.56	=C2^2	
3		3.4	-5.6	31.36	=C3^2	
4		9.8	0.8	0.64	=C4^2	
5		13.8	4.8	23.04	=C5^2	
6		17.4	8.4	70.56	=C6^2	
7	média	9	0	196.16	=SOMA(D2:D6)	

$$SQD = 196,16 \text{ cm}^2$$

Manipulação Algébrica da SQD:

$$SQD = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$SQD = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$SQD = \sum_{i=1}^n x_i^2 - \sum_{i=1}^n 2x_i\bar{x} + \sum_{i=1}^n \bar{x}^2$$

$$SQD = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2$$

$$SQD = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i \times \frac{n}{n} + n\bar{x}^2$$

$$SQD = \sum_{i=1}^n x_i^2 - 2n\bar{x} \frac{\sum_{i=1}^n x_i}{n} + n\bar{x}^2$$

$$SQD = \sum_{i=1}^n x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2$$

$$SQD = \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$SQD = \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$SQD = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

$$SQD = \sum_{i=1}^n x_i^2 - n \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2$$

$$SQD = \sum_{i=1}^n x_i^2 - n \frac{\left(\sum_{i=1}^n x_i \right)^2}{n^2}$$

$$SQD = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n}$$

Assim, temos outra fórmula para o cálculo da soma dos quadrados dos desvios

$$SQD = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$$

Essa fórmula é mais eficiente e simples do que a anterior pois não passa pelo cálculo dos desvios individuais.

No R:

```
X3<-c(0.6,3.4,9.8,13.8,17.4)
QD <- sum(X3^2) - sum(X3)^2/5
QD
```

```
#> [1] 196.16
```

```
sum(QD)
```

```
#> [1] 196.16
```

No Excel temos:

	A	B	C
1		X3	
2		0.6	
3		3.4	
4		9.8	
5		13.8	
6		17.4	
7	soma	45	=SOMA(B2:B6)
8	soma ²	601.16	=SOMAQUAD(B2:B6)
9	SQD	196.16	=B8-B7^2/5

$$SQD = 196,16 \text{ cm}^2$$

Variância amostral (s^2)

Agora podemos definir a mais importante medida de variabilidade, a variância.

Ela será definida como a soma de quadrado dos desvios, dividada pelos seus GL **graus de liberdade** (DF do inglês - *degrees of freedom*).

Ou seja:

$$s^2 = \frac{SQD}{GL}$$

Grau de liberdade é o número de observações independentes de uma estatística, após aplicada uma restrição. Ou seja, número de maneiras independentes pelas quais um sistema dinâmico pode se mover, sem violar qualquer restrição imposta a ele. Em outras palavras, o número de graus de liberdade pode ser definido como o número mínimo de coordenadas independentes que podem especificar a posição do sistema completamente.

Matematicamente, graus de liberdade é o número de dimensões do domínio de um vetor aleatório, ou essencialmente o número de componentes "livres" (**quantos componentes precisam ser conhecidos antes que o vetor seja totalmente determinado**).

Tomemos como exemplo a amostra X3:

X3
0.6
3.4
9.8
13.8
17.4

A média da amostra é dada pela soma dos valores dividido por n , ou seja, não existe qualquer restrição nesse cálculo pois, matematicamente, precisaremos de n elementos conhecidos para conhecer todo o vetor de dados $X3$.

$$\bar{x} = \frac{0,6+3,4+9,8+13,8+17,4}{n} = 9 \text{ cm}$$

Agora, tomemos o vetor de desvios de $X3$ em relação à sua média:

Desvios
-8.4
-5.6
0.8
4.8
8.4

Observe que não precisamos conhecer seus n elementos para conhecermos o vetor **Desvios**, basta conhecermos $n - 1$ elementos. Isso acontece porque sabemos, de antemão, que a soma dos desvios é igual a zero, como mostrado no slide 22:

$$\sum_{i=1}^n x_i = 0 \text{ Portanto, qualquer elemento do vetor}$$

Desvios pode ser conhecido pelas somas dos demais elementos multiplicada por (-1) , ou seja:

$$-8,4 = -1 \cdot (-5,6 + 0,8 + 4,8 + 8,4)$$

$$-5,6 = -1 \cdot (-8,4 + 0,8 + 4,8 + 8,4)$$

$$+0,8 = -1 \cdot (-8,4 - 5,6 + 4,8 + 8,4)$$

$$+4,8 = -1 \cdot (-8,4 - 5,6 + 0,8 + 8,4)$$

$$+8,4 = -1 \cdot (-8,4 - 5,6 + 0,8 + 4,8)$$

Lebrando que:

$$SQD = \sum_{i=1}^n (x_i - \bar{x})^2$$

Então essa estatística tem $n - 1$ graus de liberdade, levando o cálculo da variância amostral para:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

ou, pela segunda fórmula de SQD:

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n - 1}$$

A variância é a medida de dispersão que leva em conta todas as observações, definida como a média da soma de quadrados dos desvios (SQD) em relação à média aritmética: Para uma população, temos:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Para os dados de X3, podemos calcular a variância como:

No R:

```
X3<-c(0.6,3.4,9.8,13.8,17.4)  
var(X3)
```

```
#> [1] 49.04
```

No Excel temos:

	A	B	C	D
1		X3		
2		0.6		
3		3.4		
4		9.8		
5		13.8		
6		17.4		
7		49.04	=VAR.A(B2:B6)	
8				

$$SQD = 49,04 \text{ cm}^2$$

Desvio Padrão amostral (s)

A *variância* apresenta a unidade dos dados quadrática, o que dificulta a sua comparação. A raiz quadrada da variância é denominada **desvio padrão**: $s = \sqrt{s^2}$

Para a população temos: $\sigma = \sqrt{\sigma^2}$

A vantagem dessa estatística é que ela apresenta a mesma unidade dos dados originais.

No R:

```
X3<-c(0.6,3.4,9.8,13.8,17.4)
sd(X3)
```

```
#> [1] 7.002857
```

No Excel:

	A	B	C	D
1		X3		
2		0.6		
3		3.4		
4		9.8		
5		13.8		
6		17.4		
7		7.002857	=DESVPAD.A(B2:B6)	

$SQD = 7,003 \text{ cm}$

Erro padrão da média $[s(m)]$

Ao tomarmos r amostragens da mesma população, tendo essas amostras individuais com tamanho (n), obteremos diversas estimativas da média, distintas entre si. A partir dessas diversas estimativas da média, podemos estimar a variância, considerando os desvios de cada média, em relação à média de todas as amostras. Temos, portanto, uma estimativa da variância média, denominadas **erro padrão da média**

Essa medida fornece a ideia de precisão da estimativa da média, ou seja, quanto *menor* ela for, **maior** a precisão terá a estimativa da média. E deve ser calculada como:

$$s(m) = \frac{s}{\sqrt{n}}$$

Portanto, o aumento de n , implica na diminuição no valor de $s(m)$ e aumento da precisão da estimativa da média amostral.

```
X3<-c(0.6,3.4,9.8,13.8,17.4)
sd(X3)/length(X3)
```

```
#> [1] 1.400571
```

$s(m) = 1,400 \text{ cm}$

Coeficiente de Variação (CV)

Expressa percentualmente o desvio padrão por unidade de média, observe que a média e o desvio padrão apresentam a mesma unidade, portanto, o coeficiente de variação é um número adimensional.

$$CV = 100 \cdot \frac{s}{\bar{x}}$$

é interpretado como a variabilidade dos dados em relação à média.

Exemplo: Suponha dois grupos de indivíduos, um deles com idades $\{3, 1, 5\}$ anos e no outro, têm idades 55, 57, 53 anos. No primeiro grupo, a média de idade é 3 anos e, no segundo grupo e a $\bar{x} = 55$ *anos* anos e ambos com $s = 2$ *anos*. Onde o desvio padrão é mais importante? Por quê?

```
cv <- function(x) 100*sd(x)/mean(x)
g1<-c(3,1,5)
g2<-c(55,57,53)
cv(g1)
```

```
#> [1] 66.66667
```

```
cv(g2)
```

```
#> [1] 3.636364
```

Coeficiente de Variação (CV)

Assim, desvios de 2 anos são muito mais importantes para o primeiro grupo que para o segundo, isto é, a dispersão dos dados em torno da média é muito grande no primeiro grupo.

Deste modo, o CV pode ser usado como um índice de variabilidade, sendo que sua grande utilidade é permitir a comparação das variabilidades de diferentes conjuntos de dados.

Medidas de Forma de Distribuição

Coeficiente de Assimetria (Skewness – G1)

A interpretação da diferença entre a média e a mediana é, muitas vezes, subjetiva, então, utilizamos esse coeficiente para resumir a assimetria das observações.

Este coeficiente indica se os desvios da média são maiores para um lado da distribuição do que para o outro. É formalmente definido a partir do terceiro momento da média (OBS: a variância é o segundo momento m_2 e a média o primeiro momento m_1).

O terceiro momento pode ser computado como:

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3 \text{ assim, } G_1 = \frac{m_3}{m_2 \sqrt{m_2}} = \frac{m_3}{s^3}$$

Usualmente, o Coeficiente de Assimetria pode ser estimado pela fórmula:

$$g_1 = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

No R, vamos utilizar a função `skewness()` do pacote `agricolae` para calcular o coeficiente de assimetria para as variáveis `altura` e `idade_anos`.

```
library(agricolae)
dados_turmas %>%
  summarise(
    g1_idade = skewness(idade_anos),
    g1_altura = skewness(altura)
  )
```

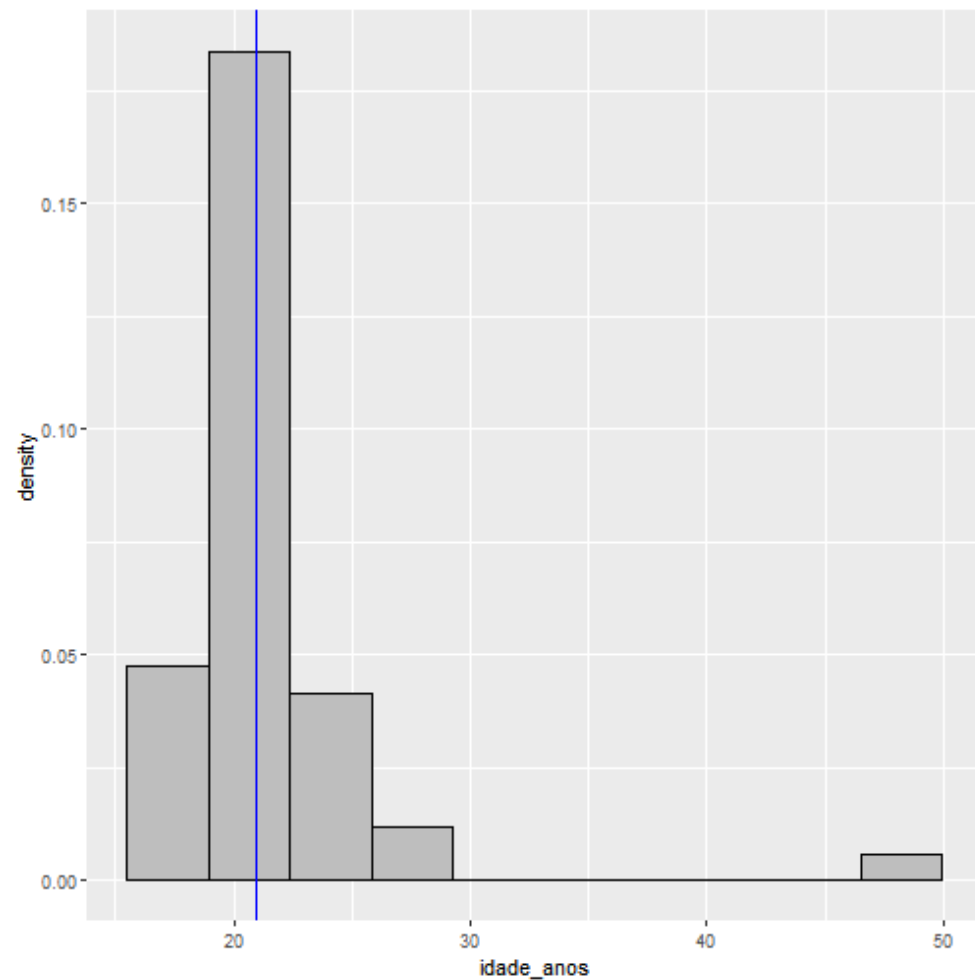
```
#> # A tibble: 1 x 2
#>   g1_idade g1_altura
#>   <dbl>    <dbl>
#> 1     4.82  -0.00283
```

- Se as observações apresentam distribuição simétrica, $g_1=0$, ou próximas a 0 (**ALTURA**).
- As observações apresentam assimetria positiva ($g_1 > 0$) se o histograma apresenta uma influência dos valores mais altos, maiores que a média e a mediana (**IDADE_ANOS**).
- As observações apresentam assimetria negativa ($g_1 < 0$) se o histograma apresenta uma influência dos valores mais baixos, menores que a média e a mediana.

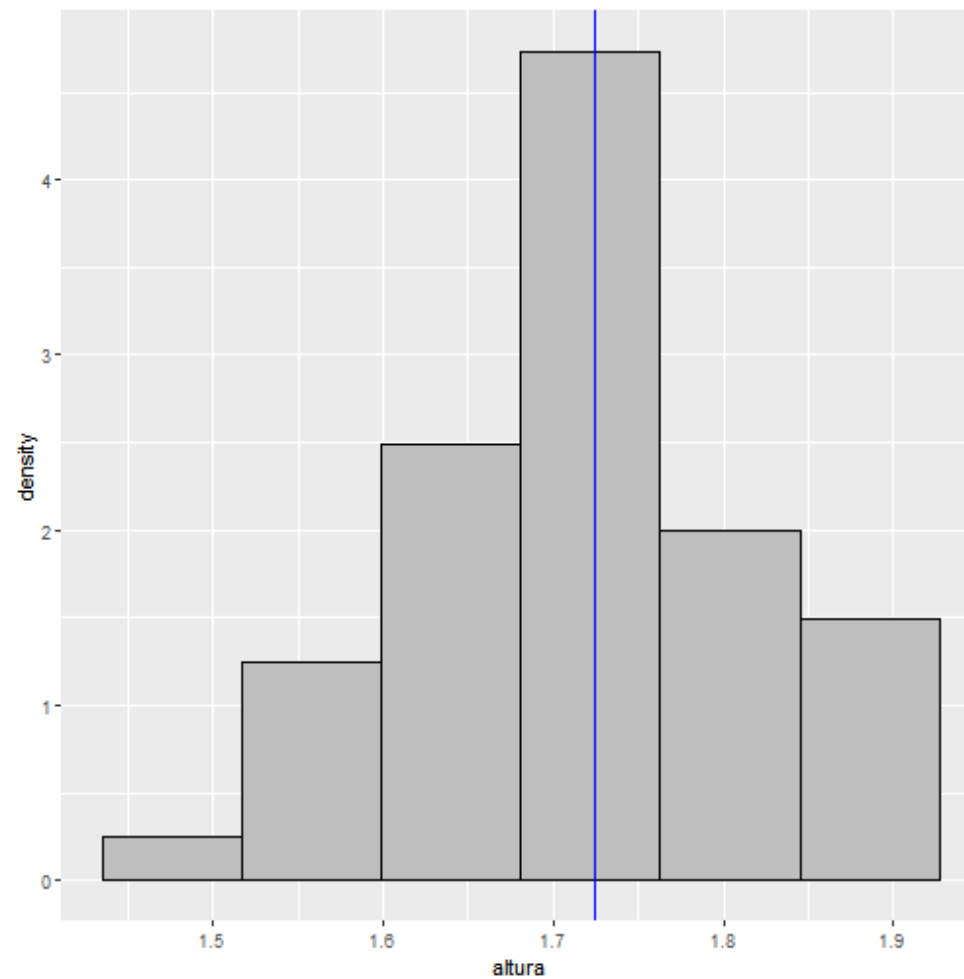
No Excel utilizamos a função =DISTRORÇÃO().

idade_anos		
18		
23		
22		
23		
23		
18		
20		
18		
20		
4.8212856	=DISTRORÇÃO(F2:F50)	

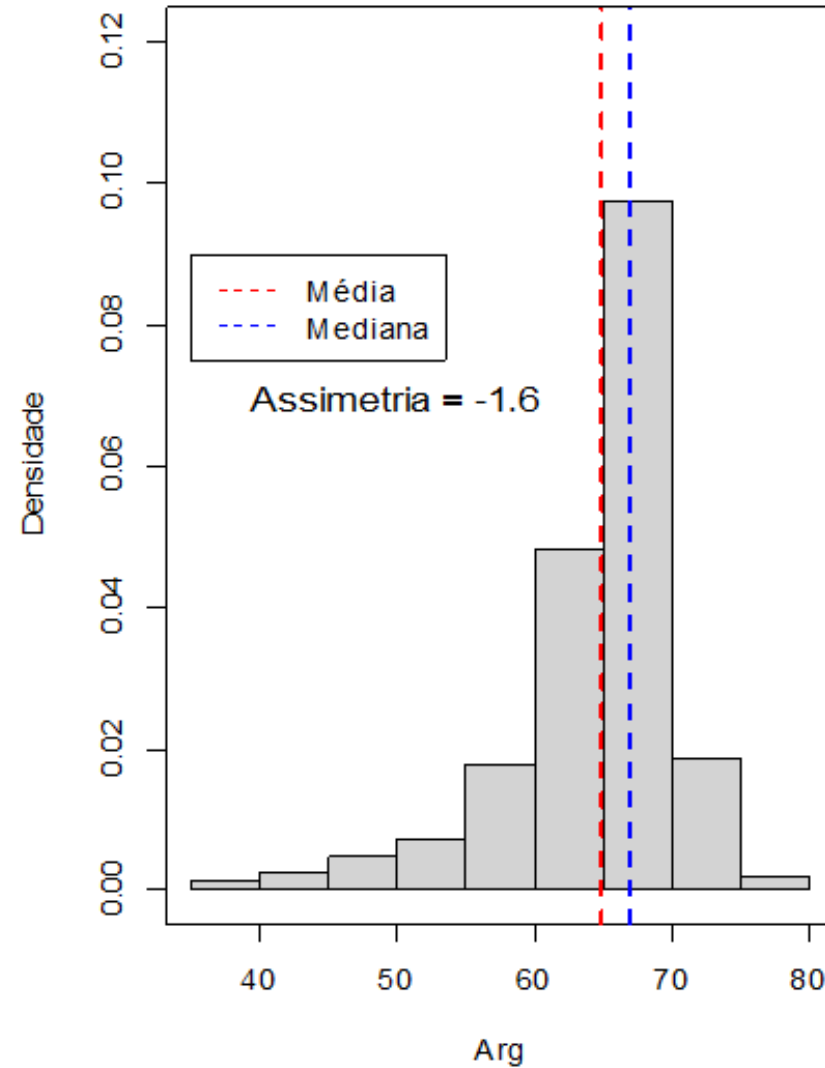
Distribuição **Assimétrica Positiva** $g_1 = 4.82$



Distribuição **Simétrica** $g_1 = -0,00283$



Abaixo segue um exemplo de distribuição **Assimétrica Negativa** da variável teor de Argila ao longo de um transecto em área de produção de cana-de-açúcar na cidade de Jaboticabal-SP.



Coeficiente de Curtose (Kurtosis – G2)

Indica o grau de achatamento de uma distribuição, é a medida do peso das caudas da distribuição. É formalmente definido a partir do quarto momento.

$$G_2 = \frac{m_4}{s_4} - 3, \text{ sendo: } m_4 = \frac{\sum_{i=1}^N (x_i - \bar{x})^4}{n}, \text{ onde } s_4 = Var(X)^2$$

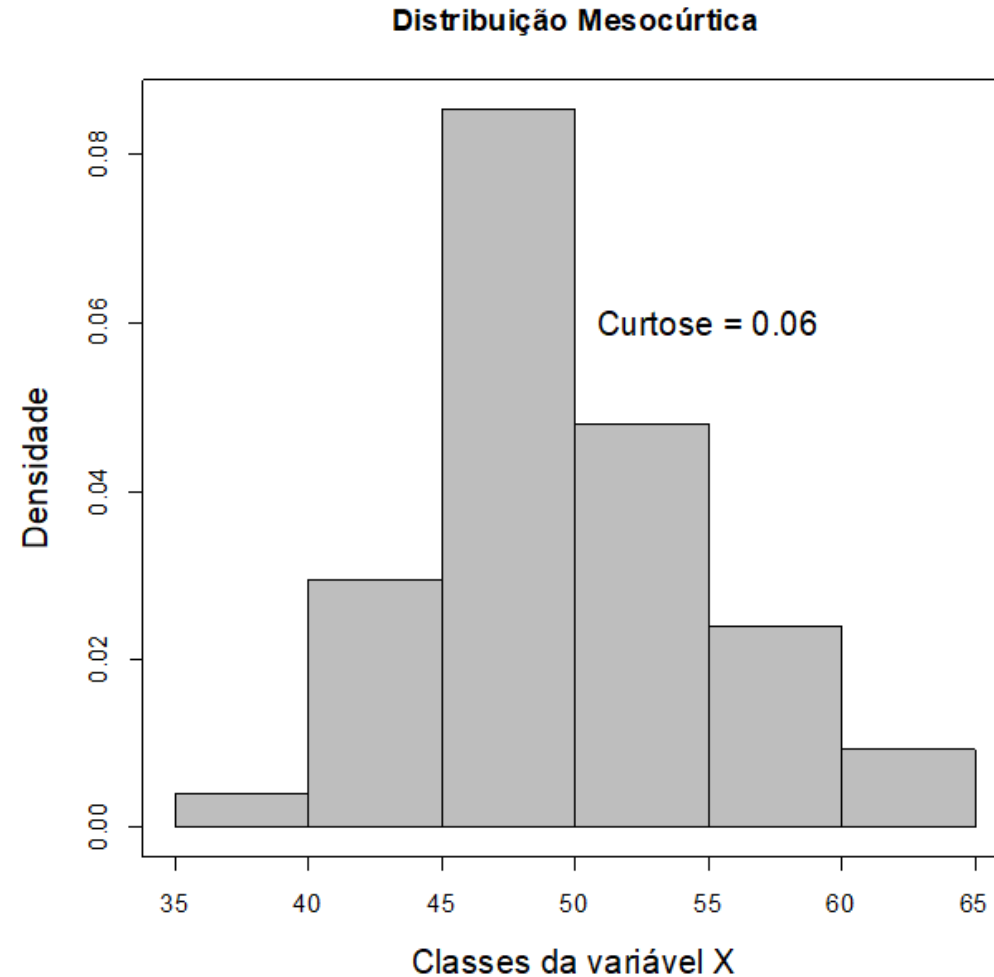
A estimativa do Coeficiente de Curtose é dada por (g_2):

$$g_2 = n(n+1) \frac{\frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4}}{(n-1)(n-2)(n-3)} - 3 \frac{(n-1)^2}{(n-2)(n-3)}$$

```
dados_turmas %>%  
  summarise(  
    g2_idade = kurtosis(idade_anos),  
    g2_altura = kurtosis(altura)  
  )
```

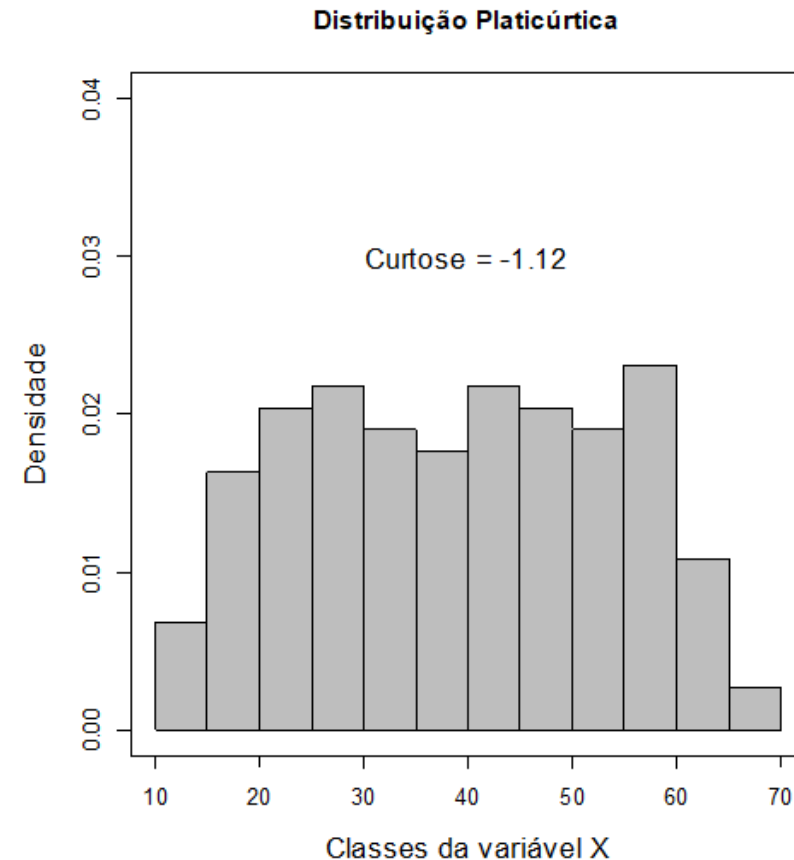
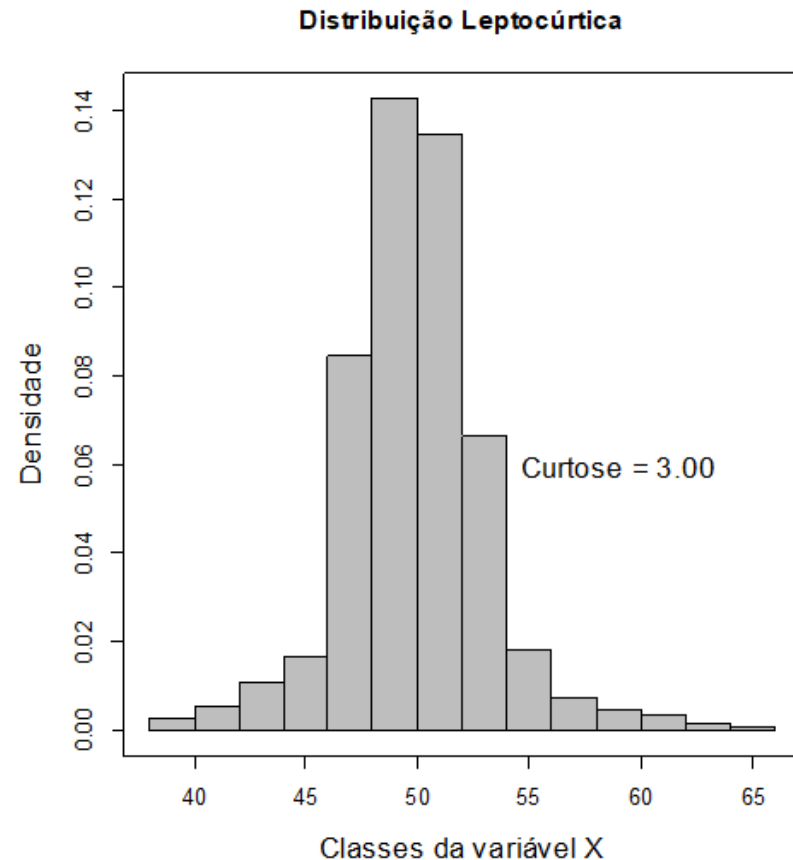
```
#> # A tibble: 1 x 2  
#>   g2_idade g2_altura
```

Se as observações seguem uma distribuição **normal**, então o coeficiente de curtose é zero: $g_2 = 0$, nesse caso a distribuição é denominada **mesocúrtica**.



Se o coeficiente de curtose é positivo, $g_2 > 0$, nesse caso o peso das caudas é baixo, distribuição é denominada **leptocúrtica**.

Se o coeficiente de curtose é negativo, $g_2 < 0$, nesse caso o peso das caudas é alto, distribuição é denominada **platicúrtica**.



No Excel, utilizamos a função =CURT().

idade_anos	
22	
23	
23	
18	
20	
18	
20	
28.099621	=CURT(F2:F50)