Estatística e Informática

Aula 04 - Medidas Estatísticas

Alan Rodrigo Panosso alan.panosso@unesp.br

Departamento de Ciências Exatas FCAV/UNESP

(21-03-2024)

SOMATÓRIA

Conjunto de dados: Nessa notação, a variável numérica de interesse (altura, idade ou peso, por exemplo) será representada pelas letras maiúsculas do nosso alfaveto latino X, Y, Z.

O conjunto de dados terá o tamanho n, que representa o número de elementos que ele contém. Em outras palavras, n representa o tamanho da amostra, o **número de observações** ou de **realizações** da variável.

Os valores específicos assumidos por tais variáveis serão representados pelas letras minúsculas x, y e z, respectivamente, seguidas de um índice i que representa a posição daquele valor específico dentro do conjunto de dados.

Assim, para distinguir um valor do outro, utilizamos esse índice i, que pode ser entendido como uma *variável auxiliar*, utilizada para contagem, que se inicia na posição 1 e termina na última posição, n, abrangendo todo o seu conjunto de dados.

Assim temos:

$$\text{Altura} : X = \{x_1, x_2, \ldots, x_n\}, \text{ com } i = 1, 2, .., n.$$

$$\text{Idade} : Y = \{y_1, y_2, \ldots, y_n\}, (i = 1, 2, .., n).$$

$$\text{Peso} : Z = \{z_1, z_2, \ldots, z_n\}, (i = 1, 2, .., n).$$

Nessa notação um valor típico da variável *Altura*, será designado por x_i e o valor final por x_n.

Somatória: Ao realizar a análise de uma variável quantitativa é necessário somar todos os seus valores.

Essa operação é frenquetemente utilizada na estatística, assim, utiliza-se uma notação compacta para representar a soma de todos os valores de uma variável de interesse.

Portanto, dado a variável X a soma de todos seus valores será notada pela letra grega *sigma* maiúscula Σ:

Dados $X = \{x_1, x_2, x_3, x_4, x_5\}$, a soma desses 5 valores:

$x_1 + x_2 + x_3 + x_4 + x_5$

será representada pela notação:

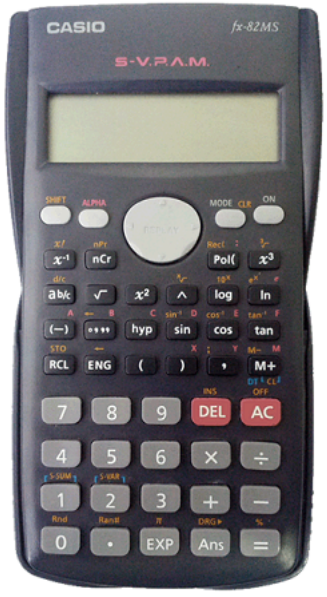
$\sum_{i=1}^{n} x_i$ ou $\sum_{i=1}^{n} x_i$

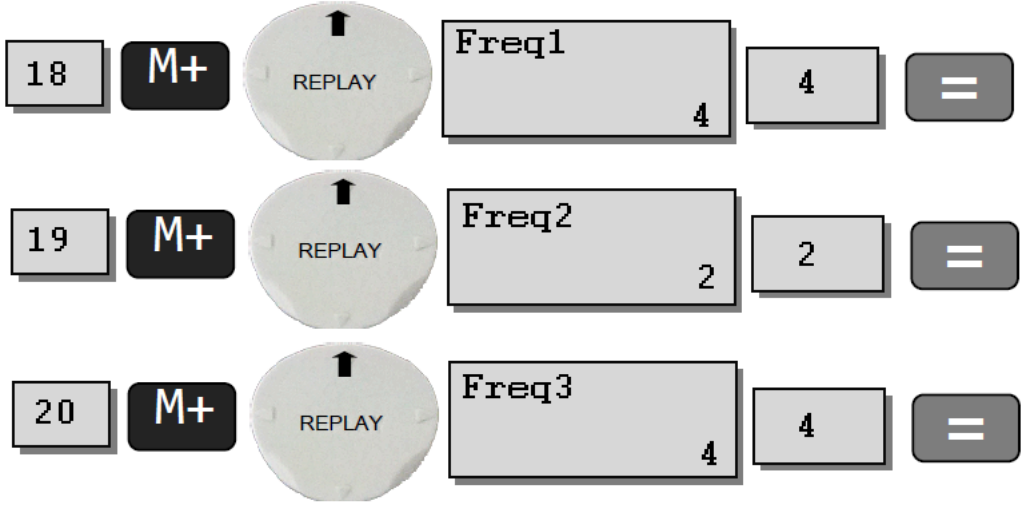
onde i atua como o índice, ou seja, a cada *iteração* ele muda e representa um dos 5 valores de X.

Utilização da Função Frequência da Calculadora Científica:

Limpe a memória da calculadora e entre com os dados $\{18, 18, 20, 20, 18, 19, 19, 20, 18, 20\}$. Organizando-os em tabela de frequência, temos:

Idade	n_i	f_i
18	4	0,4
19	2	0,2
20	4	0,4





Dado duas variáveis $X = \{3, 0, 5, 9, 7\}$ e $Y = \{2, 3, 9, 1, 2\}$, calcular:

$$\sum_{i=1}^{n} x_i = 3 + 0 + 5 + 9 + 7 = 24$$

```
X <- c(3,0,5,9,7)
sum(X)
```

Ao invés da soma ser com os índices i de 1 a n, podemos ter, por exemplo:

$$\sum_{i=2}^{4} y_i = 3 + 9 + 1 = 13$$

```
Y <- c(2,3,9,1,2)
sum(Y[2:4])
```

Observe que:

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

```
sum(X^2)
sum(X)^2
```

uma vez que:

$$3^2 + 0^2 + 5^2 + 9^2 + 7^2 \neq (3 + 0 + 5 + 9 + 7)^2$$

$$9 + 0 + 25 + 81 + 49 \neq (24)^2$$

$$164 \neq 576$$

Qual a somatória do produto entre as variáveis X e Y:

Dado $X = \{3, 0, 5, 9, 7\}$ e $Y = \{2, 3, 9, 1, 2\}$, calcular:

$$\sum_{i=1}^{n} x_i y_i$$

$$\sum_{i=1}^{n} x_i y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 + x_4 \cdot y_4 + x_5 \cdot y_5$$

$$\sum_{i=1}^{n} x_i y_i = 3 \cdot (2) + 0 \cdot (3) + 5 \cdot (9) + 9 \cdot (1) + 7 \cdot (2)$$

$$\sum_{i=1}^{n} x_i y_i = 6 + 0 + 45 + 9 + 14 = 74$$

```
sum(X*Y)
```

Propriedades da Somatória

i) A somatória de uma constante (k) é igual ao produto $n \cdot k$.

$\sum_{i=1}^{n} k = k + k + \cdots + k = n \cdot k$

ii) A somatória dos produtos de uma constante k e uma variável X é igual ao produto da constante pela soma dos valores da variável.

$$\sum_{i=1}^{n} k \cdot x_i = k \cdot x_1 + k \cdot x_2 + \cdots + k \cdot x_n = k \cdot (x_1 + x_2 + \cdots + x_n) = k \sum_{i=1}^{n} x_i$$

iii) A somatória da soma de duas variáveis é igual à adição das somatórias individuais dessas duas variáveis:

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n)$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n)$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n)$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Exercícios

1) Calcular: $\sum_{i=1}^{n} (k \cdot x_i + k) = ?$

2) Dado a média sendo:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Provar que:

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

Resposta

1) Calcular, portanto $\sum_{i=1}^{n} (k \cdot x_i + k) = ?$

$$\sum_{i=1}^{n} (k \cdot x_i + k) = \sum_{i=1}^{n} k \cdot x_i + \sum_{i=1}^{n} k$$

$$\sum_{i=1}^{n} (k \cdot x_i + k) = \sum_{i=1}^{n} k \cdot x_i + \sum_{i=1}^{n} k$$

$$\sum_{i=1}^{n} (k \cdot x_i + k) = k \cdot \sum_{i=1}^{n} x_i + n \cdot k$$

Resposta

2) Dado: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$,

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

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

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

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

Medidas de Posição
(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 da medida na reta real $(\mathfrak{R})$, ou seja, informa sobre a posição de um conjunto de dados. As principais medidas são:

1. Média:
    - **Aritmética**
    - **Ponderada**
    - Geométrica (Apostila)
    - Harmônica (Apostila)
2. Mediana
3. Moda

Média Populacional

Para a população, a média é definida como:

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

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 somatória dos valores das observações dessa variável.

A média amostral $\bar{x}$ é a estimativa mais **eficiente**, **imparcial** e **consistente** da média da população μ.

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

Onde n é o tamanho da amostra.

O exemplo a seguir apresentará o cálculo da média amostral da idade em anos de 44 alunos (`idade_anos`). Observe que a média tem a mesma unidade de medida que a das observações individuais.

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

```
# Carregando pacotes
library(tidyverse)
library(readxl)
```

Lendo o banco de dados no R

```
dados_turmas <- read_excel("data/dados_turmas.xlsx")
```

```
# Resumo rápido dos dados
glimpse(dados_turmas)
```

```
#> Rows: 44
#> Columns: 6
#> $ id          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
#> $ sexo        <chr> "F", "F", "F", "F", "M", "M", "M", "M", "M", "M", "M",
#> $ cor_cabelo  <chr> "CC", "CE", "CC", "CE", "CE", "CC", "CC", "CE", "CC",
#> $ cons_alcool <dbl> 4, 1, 2, 3, 4, 2, 1, 2, 2, 4, 3, 3, 3, 2, 3, 2, 3, 3,
#> $ altura      <dbl> 1.68, 1.59, 1.70, 1.50, 1.76, 1.60, 1.84, 1.88, 1.90,
#> $ idade_anos  <dbl> 19, 20, 49, 20, 23, 28, 19, 20, 20, 19, 21, 21, 21, 18
```

Cálculo da Média Aritmética

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{19 + 20 + \cdots + 23}{44} = 21,04545 \text{ anos}$$

```
dados_turmas %>%
  summarise(
    media = mean(idade_anos)
  )
```

media
21.04545

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

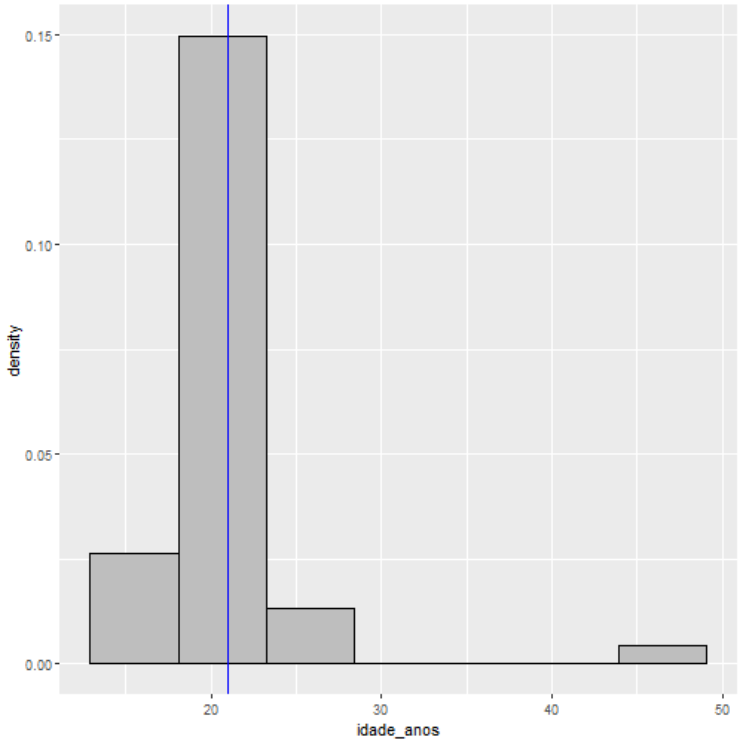
$$\bar{x} = 18 \cdot (0,1363) + 19 \cdot (0,3181818) + \cdots + 49(0,0227273)$$

$$\bar{x} = 21,04545 \; anos$$

Tabela de frequência para idade

idade_anos	ni	fi	perc
18	6	0.1363636	13.636364
19	14	0.3181818	31.818182
20	9	0.2045455	20.454545
21	3	0.0681818	6.818182
22	3	0.0681818	6.818182
23	5	0.1136364	11.363636
24	1	0.0227273	2.272727
27	1	0.0227273	2.272727
28	1	0.0227273	2.272727
49	1	0.0227273	2.272727

Histograma e média para idade_anos



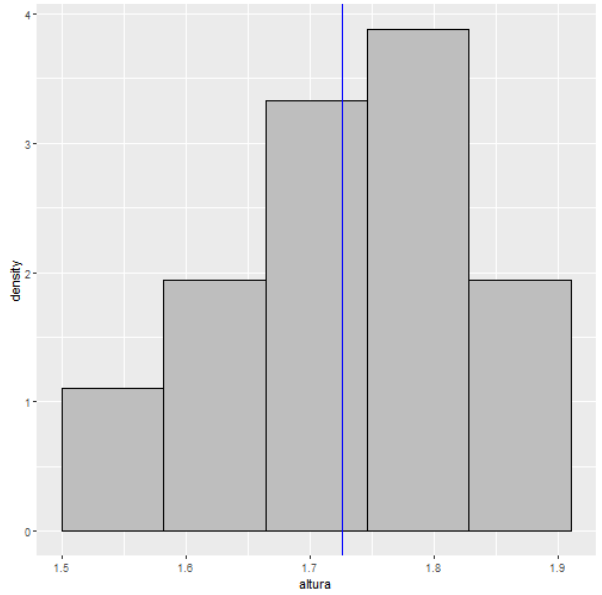
Se os dados são plotados como um histograma a média é o centro de gravidade do gráfico (linha azul).

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

Histograma e média para altura

```
dados_turmas %>%
  pull(altura) %>%
  mean()
```

```
#> [1] 1.725682
```



$$\bar{x} = \frac{1,68 + 1,59 + \cdots + 1,80}{44} = 1,7257 \text{ m}$$

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

Média Ponderada

Em algumas situações as observações têm graus de importância diferentes. Usa-se, então, a média ponderada:

$$\bar{x}_P = \frac{\sum_{i=1}^{n} x_i \cdot \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

Onde λ_i é o peso associado à i-ésima observação. Assim, ele mede a importância relativa dessa i-ésima observação em relação às demais.

Exemplo: Calcular a intensidade média de infestação do complexo "Broca-podridão" da cana-de-açúcar, larva de lepidóptera (*Diatraea saccharalis*) em plantas jovens associado à infecção por fungos, causando a podridão vermelha (*Colletotrichum falcatum*), em 8 variedades plantadas em uma propriedade rural.

Variedades	Nº de Talhões Infestados	% de Infestação
CB-40-13	12	9,10
CB 41-76	40	14,57
CB 46-47	4	3,20

$$\bar{x}_P = \frac{\sum_{i=1}^{n} x_i \cdot \lambda_i}{\sum_{i=1}^{n} \lambda_i} = \frac{12 \cdot (9{,}10) + 40 \cdot (14{,}57) + 4 \cdot (3{,}20)}{12 + 40 + 4} = 12,58\%$$

No R:

```
pesos<-c(12,40,4)
X <- c(9.1,14.57,3.2)
weighted.mean(X, pesos)
```

```
#> [1] 12.58571
```

Mediana

É o valor que ocupa a posição central do conjunto de dados ordenados (X' ou X_o) que pode ser denotada por Md.

Assim, antes de encontrar ou calcular a mediana as observações são ordenadas do menor para o maior valor.

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*:

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

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

Exemplo 1:

Dado $X = \{50, 60, 20, 50, 30, 90, 70\}$, teremos a sua versão ordenada X_0 dada por:

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

$n = 7$, ímpar.

$$Md = Xo_{\left(\frac{n+1}{2}\right)} = Xo_{\left(\frac{7+1}{2}\right)} = Xo_4 = 50$$

No R:

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

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

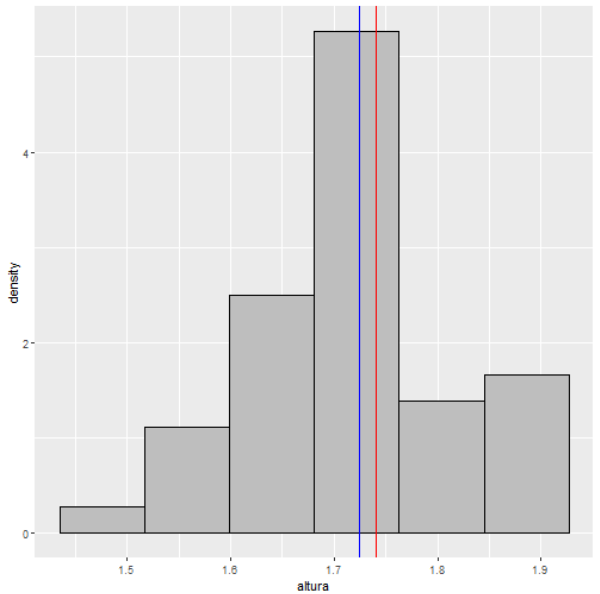
Observe que a unidade da mediana é a mesma dos dados originais.

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 $\bar{x}$.

```
dados_turmas %>%
  pull(altura) %>%
  median()
```

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



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:

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

A média será dada por:

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

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

$$S = 5, 2, 3, 6, 100, 9$$

$n = 6$, par.

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

Assim, para os dados de salários temos:

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

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

A média será dada por:

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

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

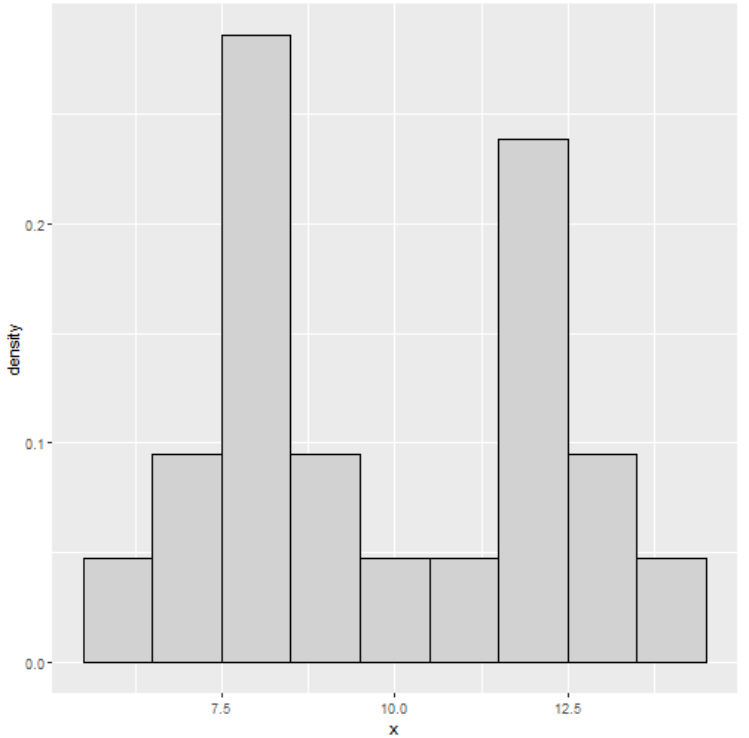
Moda

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.



Medidas Separatrizes

(Quantis)

QUANTIS

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

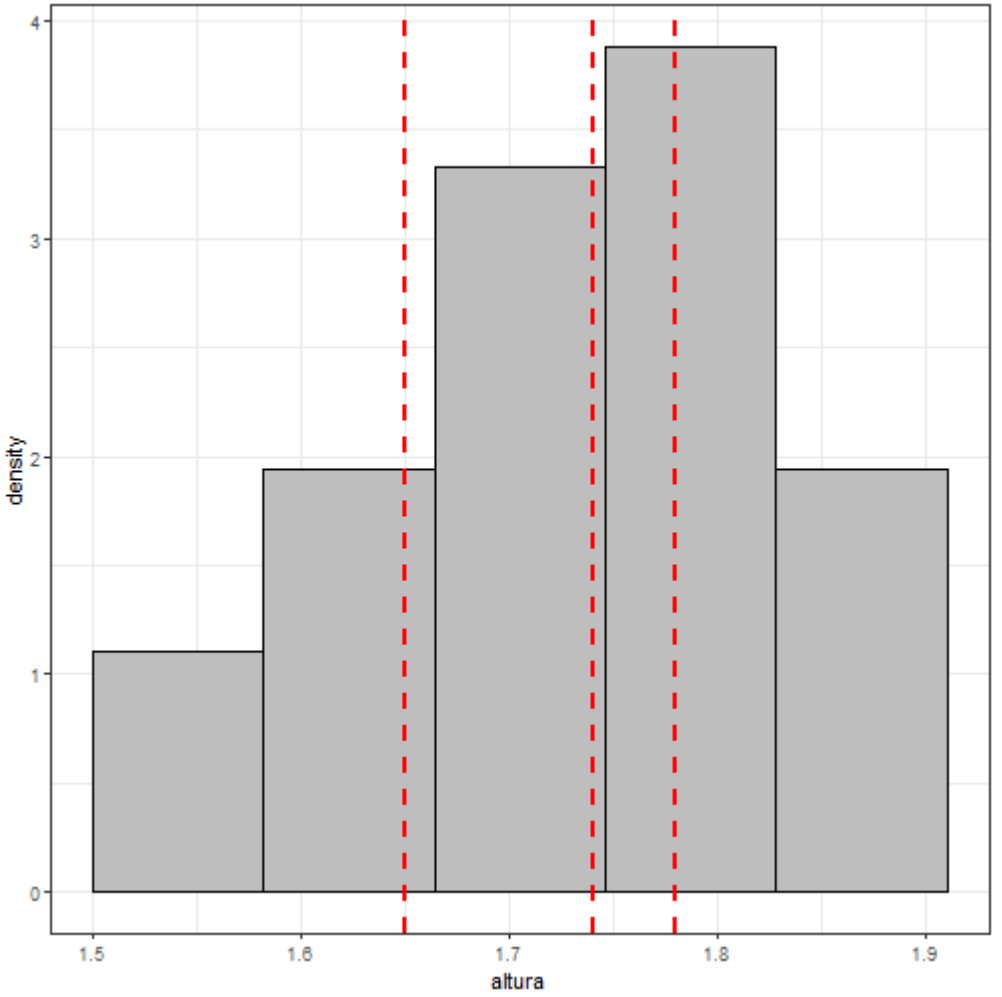
- Os Quantis de ordem $25, 50, 75$ são chamados **Quartis** (Q1, Q2, Q3). Naturalmente, Q2 = mediana (Md).
- Os **Decis** são os quantis de ordem $10, 20, ..., 90$ (D1, D2, ..., D9)
- Os **Percentis** são os quantis de ordem $1, 2, ..., 99$ (P1, P2, ..., P99).

No R, ele 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 desejado.

```
dados_turmas %>%
  summarise(
    q1=quantile(altura, 0.25),
    q2=quantile(altura, 0.50),
    q3=quantile(altura, 0.75),
  )
```

```
#> # A tibble: 1 × 3
#>      q1    q2    q3
#>   <dbl> <dbl> <dbl>
#> 1  1.67  1.74  1.76
```

Histograma da variável altura de alunos



Medidas de Dispersão (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 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

media_X1	media_X2	media_X3	media_X4
9	9	9	9

Tabela de Amplitudes

$(\Delta = Máximo - Mínimo)$

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 as 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 somatória dos erros e_i é sempre igual a zero:

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

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

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

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

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

Soma dos Quadrados dos Desvios (SDQ)

Para evitar a inconveniência da somatória 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, também, a unidade dos dados ao quadrado e pode ser calculada no R:

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

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

$SQD = 196,16\ cm^2$

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:

```
n <- length(X3)
sum(sum(X3^2) - sum(X3)^2/n)
```

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

$SQD = 196,16\ cm^2$

Variância amostral $\left(s^2\right)$

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

Ela será a soma de quadrado dos desvios, dividida pelos seus **graus de liberdade (GL)** (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 $X3$.

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

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

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

Portanto, qualquer elemento do vetor **Desvios** pode ser conhecido pelas somas dos demais elementos multiplicada por (-1), ou seja:

$$
\begin{aligned}
-8,4 &= -(-5,6 + 0,8 + 4,8 + 8,4) \\
-5,6 &= -(-8,4 + 0,8 + 4,8 + 8,4) \\
+0,8 &= -(-8,4 - 5,6 + 4,8 + 8,4) \\
+4,8 &= -(-8,4 - 5,6 + 0,8 + 8,4) \\
+8,4 &= -(-8,4 - 5,6 + 0,8 + 4,8)
\end{aligned}
$$

Lembrando que:

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

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

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

ou, pela segunda fórmula da SQD

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

A variância é a medida de dispersão que leva em conta todos 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:

```
var(X3)
```

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

$SQD = 49,04\ cm^2$

Desvio Padrão

A *variância* apresenta a unidade dos dados quadrática, o que dificulta a sua comparação e interpretação. Portanto, podemos tomar a raiz quadrada da variância, que é denominada **desvio padrão amostral** (quando calculado a partir da variância amostral):

$$s = \sqrt{s^2}$$

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

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

No R:

```
sd(X3)
```

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

Erro padrão da média $\left[s\left(m\right)\right]$

Ao tomarmos uma amostragens da mesma população, tendo essas amostras individuais com tamanho $\left(n\right)$, 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.

```
sd(X3)/sqrt(length(X3))
```

```
#> [1] 3.131773
```

$s(m) = 3,132\ cm$

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 $\bar{x} = 55\ anos$ ambos com $s = 2\ anos$. Onde o desvio padrão é mais importante? Por quê?

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

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

Para Grupo 1

$$CV = 100 \cdot \frac{2}{3} = 66,67\%$$

Para Grupo 2

$$CV = 100 \cdot \frac{2}{55} = 3,63\%$$

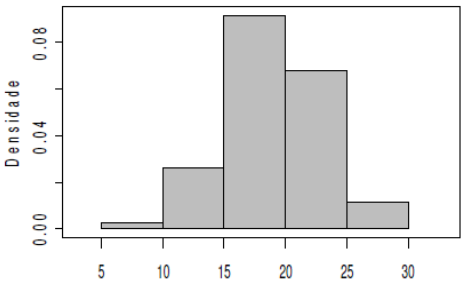
Assim, desvios de 2 anos são muito mais importantes para o primeiro grupo que para o segundo, isto é, a dispersão dos dados em torno da média é muito grande no primeiro grupo.

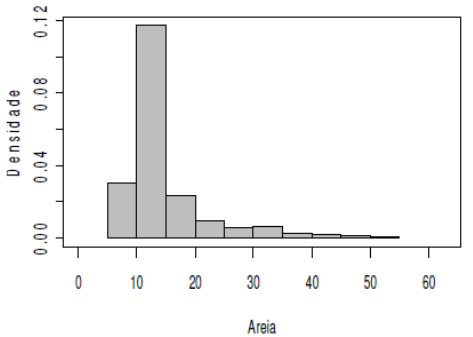
Deste modo, o CV pode ser usado como um índice de variabilidade, sendo que sua grande utilidade é permitir a comparação das variabilidades de diferentes conjuntos de dados, e variáveis.

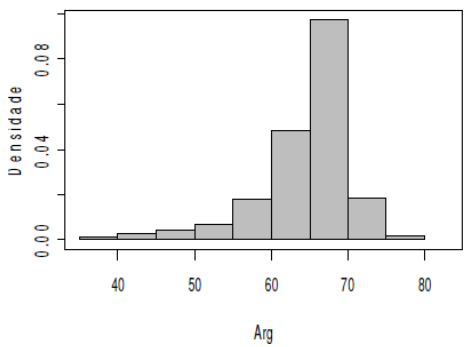
Medidas de Forma de Distribuição

Introdução

Observe a distribuição de frequência das variáveis teores de Areia, Silte e Argila (%) coletados em uma área experimental de Latossolo, utilizado na produção de cana-de-açúcar na região de Jaboticabal-SP:

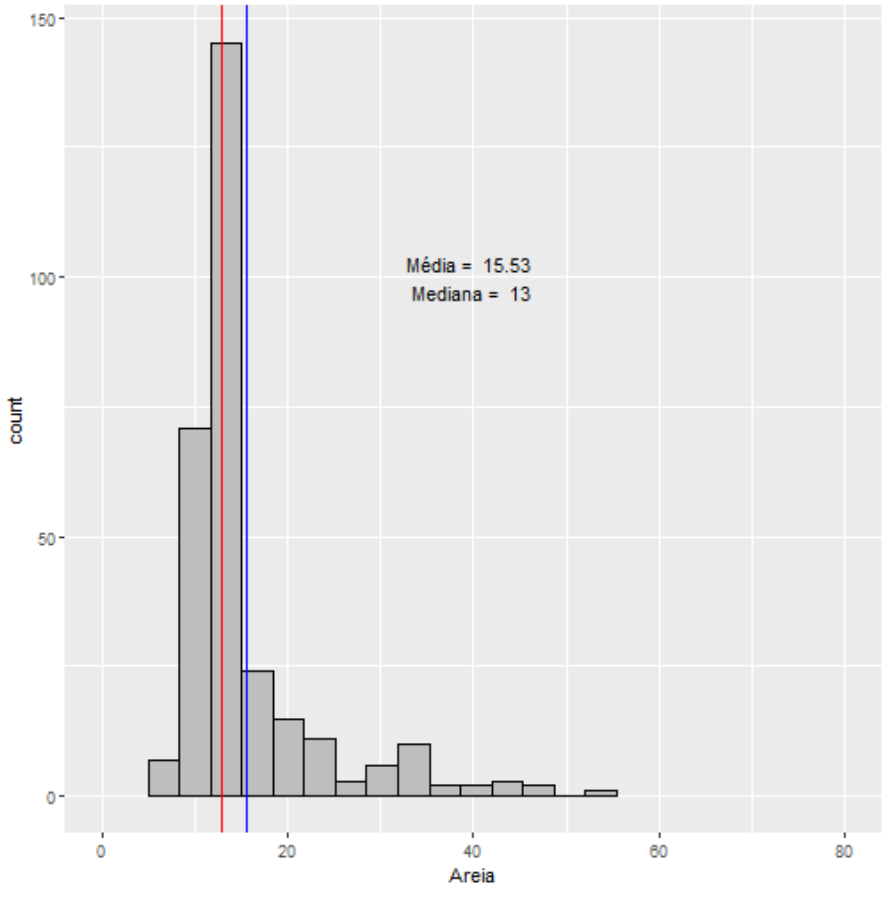






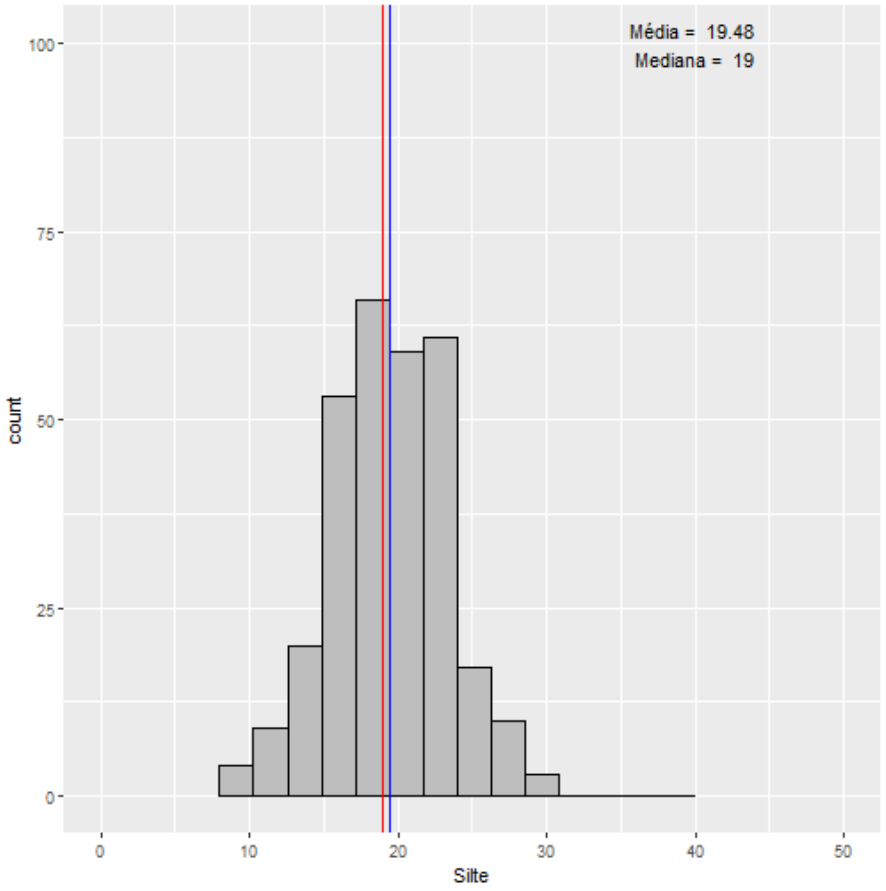
As Estatísticas Descritivas desses atributos são:

	Mediana	Média	Média - Mediana	Comparação
Silte	19	19.48	0.48	Média semelhante à mediana
Areia	13	15.00	2	Média > Mediana
Argila	67	64.98	-2.02	Média < Mediana



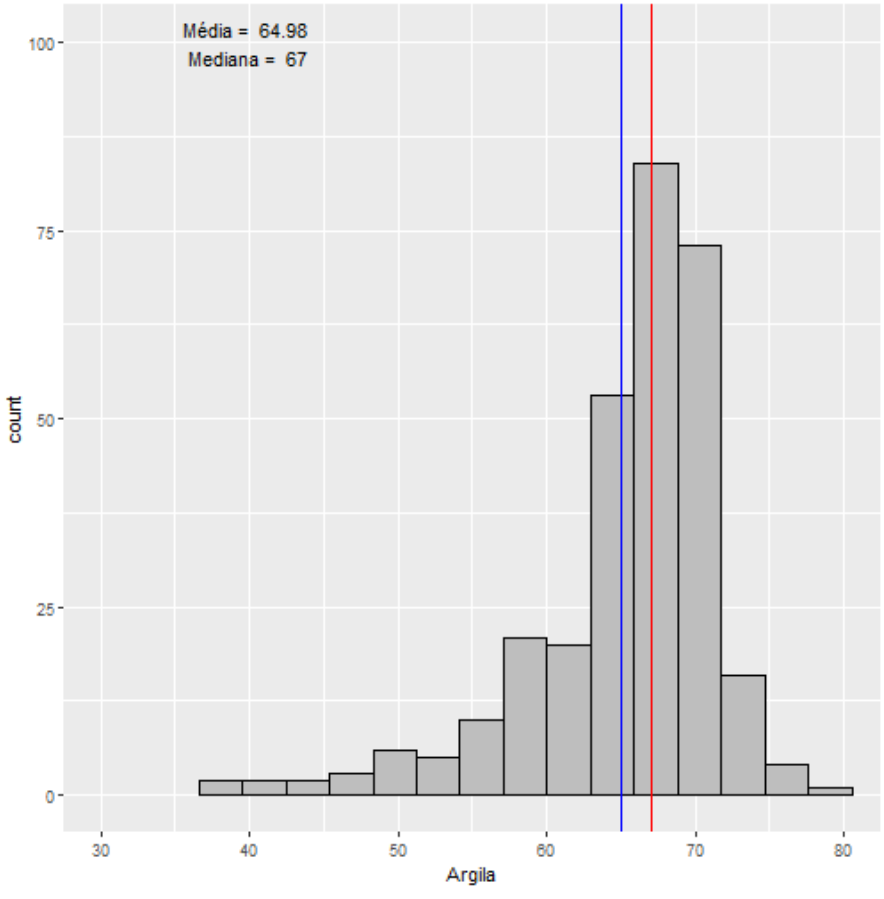
Para a **Areia**, a *Média* foi maior do que a *Mediana*, uma vez que a distribuição de frequência dessa variável indicou **uma maior concentração de observações nas classes de valores mais baixos** de areia.

Nesse caso dizemos que a distribuição de frequência é **Assimétrica Positiva**, ou seja, a média está sendo influenciada por amostras com altos valores de areia.



Observe que para o teor de **Silte** os valores de **Média** e **Mediana** foram semelhantes, existe um "pequena" diferença entre eles.

Nesses casos dizemos que a distribuição de frequência dos dados é **Simétrica**, ou seja, observações discrepantes (valores altos, ou baixos), não foram observadas.



Para o teor de **Argila**, a *Média* foi menor do que a *Mediana*, uma vez que a distribuição de frequência dessa variável indicou **uma maior concentração de observações nas classes de valores mais altos** de argila.

Nesse caso dizemos que a distribuição de frequência é **Assimétrica Negativa**, ou seja, a média está sendo influenciada por amostras com baixos valores de argila.

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 *m2* e a média o primeiro momento *m1*).

O terceiro momento pode ser computado como:

$m_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^3$

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:

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

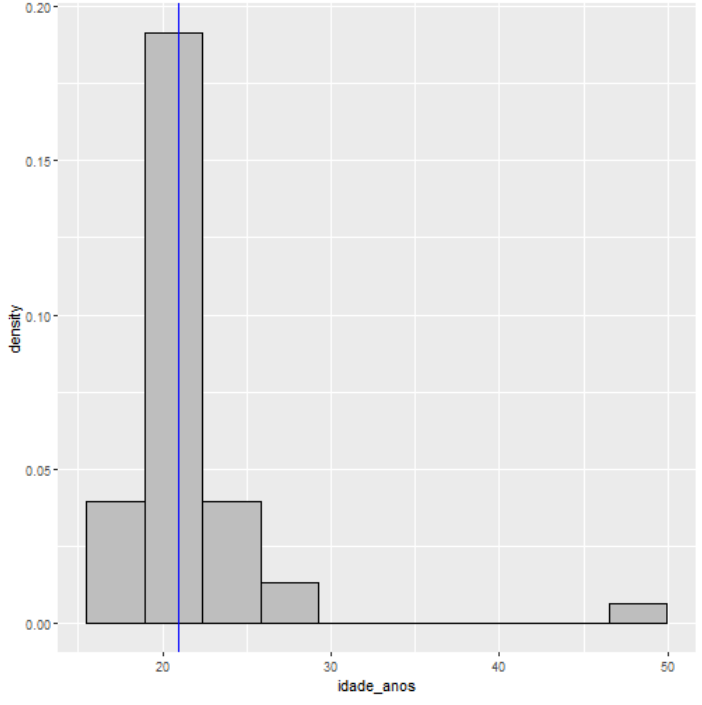
No R, vamos utilizar a função `skewness()` do pacote `{agricolae}` para calcular o coeficiente de assimetria para as variáveis `altura` e `idade_anos`.

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

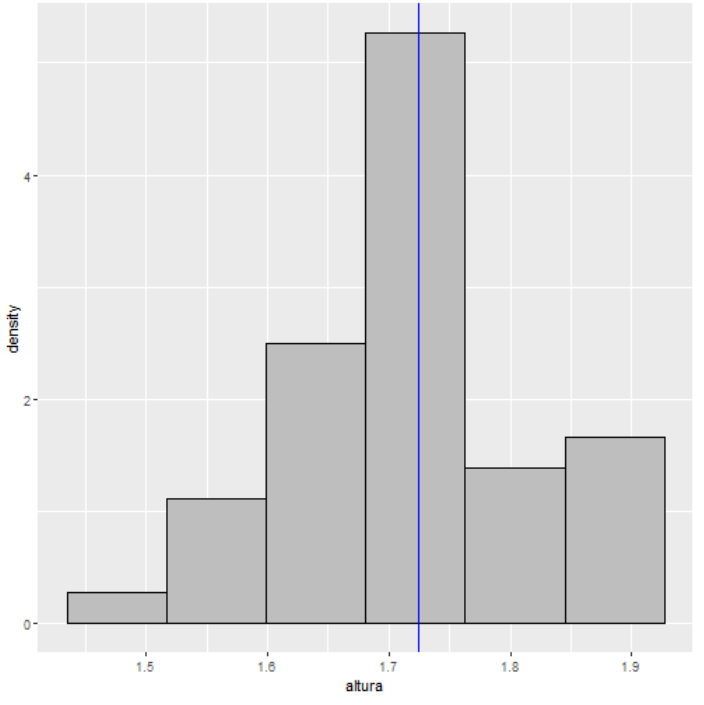
```
#> # A tibble: 1 × 2
#>   g1_idade g1_altura
#>      <dbl>     <dbl>
#> 1     4.67    0.0468
```

- 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.

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



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



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):

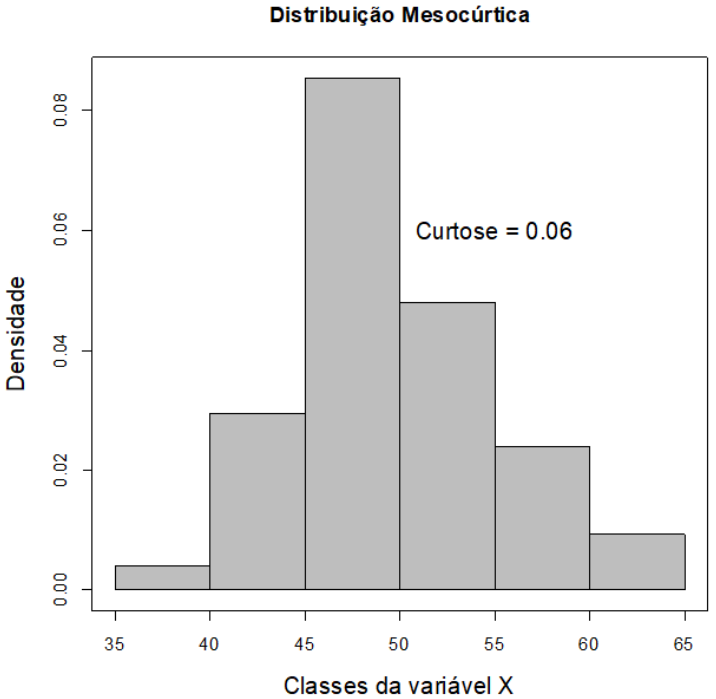
$$g_2 = n(n+1)\frac{\frac{\sum_{i=1}^{n}(x_1-\bar{x})^4}{s^4}}{(n-1)(n-2)(n-3)} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

No R podemos calcular o coeficiente de curtose a partir da função `kurtosis` do pacote `{agricolae}`

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

```
#> # A tibble: 1 × 2
#>   g2_idade g2_altura
#>      <dbl>     <dbl>
#> 1     26.0    -0.170
```

Se as observações seguem uma distribuição **normal**, o próximas à normal, então o coeficiente de curtose é próximo a 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**.