Medidas descritivas

- ***Medidas descritivas:***
  - ***Objectivo***: Sumariar as características das amostras.
  - ***Ex***:
    - Medidas de localização: Média, Moda, Mediana, Quantis, etc.
    - Medidas de dispersão: Variância, Desvio padrão, Coeficiente de variação, etc.
    - Medidas de forma: Coeficiente de assimetria, Coeficiente de Kurtose, etc.

  - ***Def*:** A ***média*** (aritmética), $\bar{x}$, de uma amostra de n observações não agrupadas, x_1, x_2, ..., x_n, é dada por:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Ex: Calcule a média das observações {1,1,3,5,7}.

$$\bar{x} = (1+1+3+5+7)/5 = 3.4$$

$R: mean()$

Medidas descritivas

- ***Def***: A ***média***, $\overline{x}$, de uma amostra de n observações agrupadas em c classes é dada por:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{c} x_i n_i = \sum_{i=1}^{c} x_i f_i$$

Nota: Para dados contínuos, x_i representa a marca da classe i.

Ex: Calcule a média das observações {1,1,3,5,7}, agrupando as classes previamente.

$$\overline{x} = (1\times 2 + 3\times 1 + 5\times 1 + 7\times 1)/5 = 3.4$$

- ***Def***: A ***moda***, *Mo*, de dados discretos ou contínuos não agrupados é o valor observado com maior frequência.

Ex: Para as observações {1,1,3,5,7}, Mo = 1.

- ***Def***: A ***classe modal*** de um conjunto de dados contínuos agrupados é a classe com maior frequência.

Medidas descritivas

- ***Def*:** A ***mediana***, *Me*, duma série de observações $x_1, x_2, \ldots, x_n$, é o valor que divide a série em 2 partes iguais, tal que 50% das observações têm valor menor ou igual a *Me*. Calcula-se da seguinte forma:

a) Se os dados são discretos ou contínuos não agrupados e as observações estão ordenadas por ordem crescente, então:

$$Me = \begin{cases} \dfrac{x_{n/2} + x_{n/2+1}}{2} & , n \text{ par} \\ x_{(n+1)/2} & , n \text{ impar} \end{cases}$$

$R: median()$

Ex: Para {1,1,3,5,7}, $n = 5$ e $(n+1)/2 = 3$. Logo, $Me = x_3 = 3$.

b) Se os dados são contínuos agrupados, então:

$$Me = L_k + \frac{h_k}{f_k}\left(0.5 - F_{k-1}\right)$$

onde:

$F_{k\text{-}1}$: Frequência relativa acumulada até à classe $k - 1$.

k: Classe mediana, tal que $F_{k\text{-}1} < 0.5$ e $F_k \geq 0.5$.

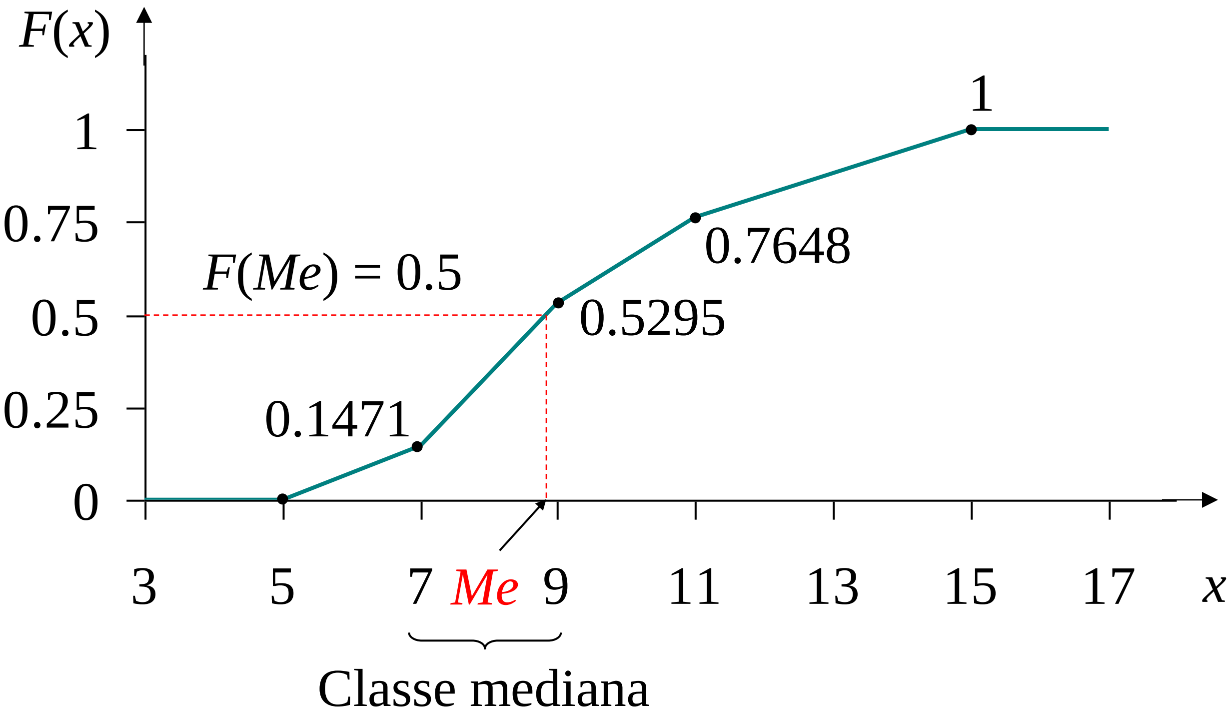
L_k: Limite inferior da classe mediana.

h_k: Amplitude da classe mediana.

f_k: Frequência relativa da classe mediana.

Ex:

Classe	n_i	f_i	F_i
[5, 7[	5	0.1471	0.1471
[7, 9[	13	0.3824	0.5295
[9, 11[	8	0.2353	0.7648
[11, 15[	8	0.2353	1.0001



<u>Por interpolação linear</u>:

$$y = \frac{0.5295 - 0.1471}{9 - 7}(x - 7) + 0.1471 \quad \Leftrightarrow \quad x = \overset{L_2}{7} + \frac{\overset{h_2}{2}}{\underset{f_2}{0.3824}}(y - \overset{F_1}{0.1471})$$

$$\therefore \quad Me = x(0.5) \approx 8.8457$$

Medidas descritivas

- ***Def***: O ***quantil de ordem α***, z_α, $(0<\alpha<1)$, duma série de observações x_1, x_2, ..., x_n é calculado da seguinte forma:

a) Se os <u>dados são discretos ou contínuos não agrupados</u> e as observações estão ordenadas por ordem crescente, então:

$$z_\alpha = x_k$$

$R: quantile(\ldots, type = 1)$

onde k é o maior inteiro menor que $n\alpha + 1$

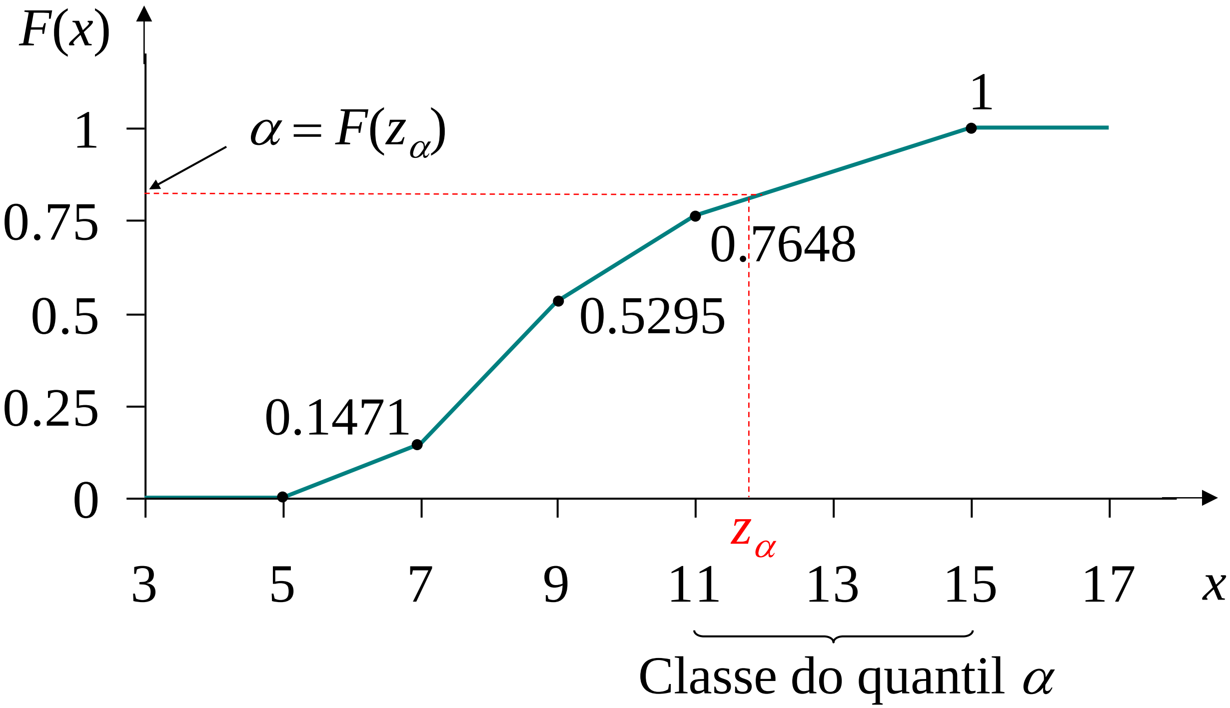
Ex: Seja {1,1,3,5,7} a amostra e $\alpha = 0.1$. Então, $n = 5$ e $n\alpha + 1 = 1.5$. Logo, $z_{0.1} = 1$.

b) Se os <u>dados são contínuos agrupados</u>, z_α calcula-se como a mediana mas substituindo a classe mediana pela *classe do quantil* α e o valor 0.5 por α.

$$z_\alpha = L_k + \frac{h_k}{f_k}\left(\alpha - F_{k-1}\right)$$

k: Classe do quantil α, tal que $F_{k-1} < \alpha$ e $F_k \geq \alpha$

Ex:



$F(x)$
$\alpha = F(z_\alpha)$
1
0.75
0.5
0.25
0
1
0.7648
0.5295
0.1471
z_α
3
5
7
9
11
13
15
17
x
Classe do quantil α

Medidas descritivas

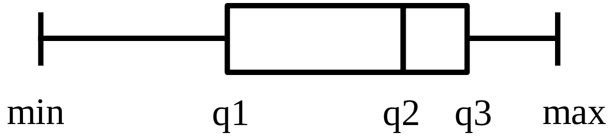
- ***Def*: O *percentil de ordem k*, p_k, (k=1,2,...,99), de um conjunto de dados é o quantil de ordem k/100. Logo, k % das observações são menores do que p_k.**

- ***Def*: O *decil de ordem k*, d_k, (k=1,2,…,9), de um conjunto de dados é o quantil de ordem k/10.**

- ***Def*: O *quartil de ordem k*, q_k, (k=1,2,3), de um conjunto de dados é o quantil de ordem k/4.**

- *Nota*: Me = $p_{50} = d_5 = q_2$

- ***Def*: Um *diagrama de caixa* (*box plot*) é uma representação gráfica de dados baseada nos seus quartis que mostra a dispersão desses dados.**



R: $boxplot()$

- ***Def***: A ***variância***, s^2, de uma amostra de n observações não agrupadas, x_1, x_2, ..., x_n, é dada por:

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{k=1}^{n} x_k^2 - n\bar{x}^2\right] = \frac{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}{n(n-1)}$$

$R: var()$

- ***Def***: A ***variância***, s^2, de uma amostra de n observações agrupadas em c classes é dada por:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{c}(x_i - \bar{x})^2 n_i = \frac{1}{n-1}\left[\sum_{i=1}^{c} x_i^2 n_i - n\bar{x}^2\right] = \frac{n\sum_{i=1}^{c} x_i^2 n_i - \left(\sum_{i=1}^{c} x_i n_i\right)^2}{n(n-1)}$$

- ***Def***: O ***desvio padrão***, s, de uma amostra é a raiz quadrada da sua variância.

$R: sd()$

Ex: Calcule o desvio padrão das observações {4,9,11,14,22}.

$$s^2 = \frac{5\left(4^2+9^2+11^2+14^2+22^2\right)-(4+9+11+14+22)^2}{5\times 4} = 44.5 \;\therefore\; s = \sqrt{44.5} \approx 6.67$$

Medidas descritivas

- ***Def*: O *coeficiente de variação* duma amostra é dado por:**

$$CV = \frac{s}{|\bar{x}|}$$

- Vantagem: O CV é uma medida adimensional.
- Desvantagem: Não está definido para $\bar{x} = 0$. Logo, só deve ser usado quando as observações têm todas o mesmo sinal.

Ex: Calcule o coeficiente de variação das observações {4,9,11,14,22}.

$$\bar{x} = (4+9+11+14+22)/5 = 12$$

$$CV = \frac{6.67}{12} \approx 0.556 = 55.6\%$$

Medidas descritivas

- ***Def*:** O ***momento amostral <u>centrado</u> de ordem r***, ($r \in \mathbb{N}_0$), é calculado da seguinte forma:

a) Para <u>*n* observações não agrupadas</u>:

$$m_r = \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^r$$

b) Para <u>*n* observações agrupadas em *c* classes</u>:

$$m_r = \frac{1}{n}\sum_{i=1}^{c}\left(x_i - \bar{x}\right)^r n_i$$

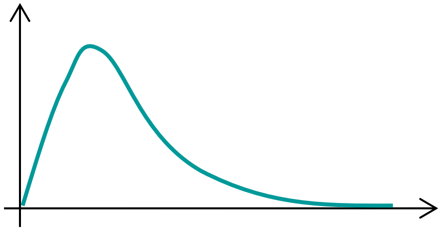
Medidas descritivas

- ***Def*: O *coeficiente de assimetria amostral* , a_3 ,** é dado por

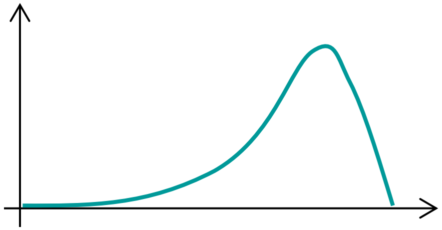
$$a_3 = \frac{m_3}{s^3}$$

- *Notas*:
  - É adimensional
  - Mede a assimetria da distribuição
  - $a_3 = 0$ quando é simétrica
  - $a_3 > 0$ quando a cauda direita é a mais comprida
  - $a_3 < 0$ quando a cauda esquerda é a mais comprida

***Ex*:**



Distribuição enviesada à direita; assimétrica positiva



Distribuição enviesada à esquerda; assimétrica negativa

Medidas descritivas

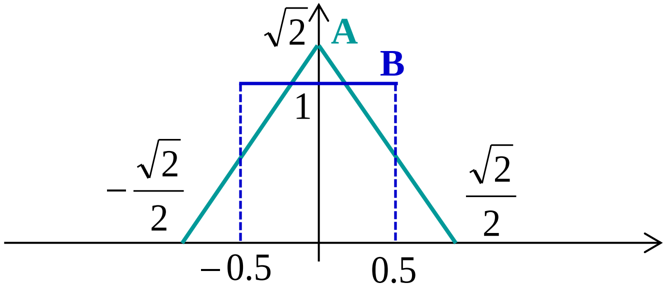
- ***Def***: O ***coeficiente de curtose amostral*** **,** $\boldsymbol{a_4}$ **,** é dado por

$$a_4 = \frac{m_4}{s^4}$$

- *Notas*:

  - É adimensional
  - Mede o achatamento e o peso das caudas da distribuição
  - $a_4 = 3 \Rightarrow$ Mesmo achatamento que a normal (*mesocúrtica*)
  - $a_4 > 3 \Rightarrow$ Mais esguia e cauda mais pesada que a normal (*leptocúrtica*)
  - $a_4 < 3 \Rightarrow$ Mais achatada e cauda menos pesada que a normal (*platicúrtica*)

Ex:



a_4(A) > a_4(B) porque B é mais achatada e tem caudas menos pesadas