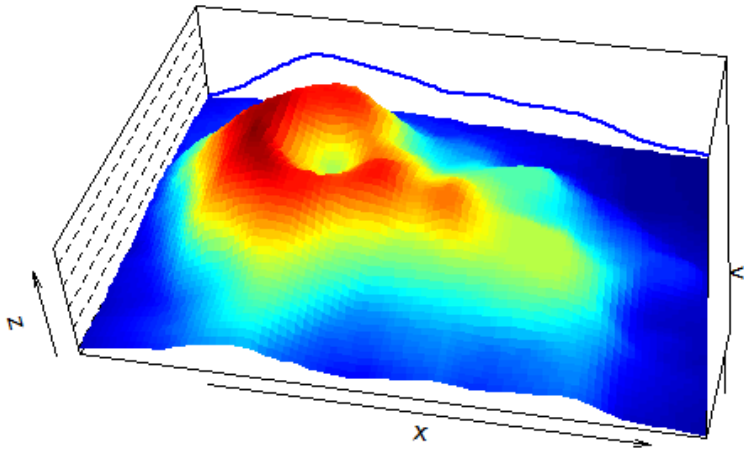


Estatística



Profa. Dra. Juliana Garcia Cespedes
PEDS - Programa de Excelência em Data Science
ITA+ITAÚ



Medidas resumo RESISTENTES

Quantis / Percentil

- Apenas com os valores da média e do desvio padrão, não é possível identificar o comportamento da distribuição (assimétrico ou simétrico).
- Pode-se definir uma medida, chamada **quantil de ordem p** , indicada por **$q(p)$** , onde p é uma proporção qualquer, $0 < p < 1$, tal que $100p\%$ das observações sejam menores do que $q(p)$.

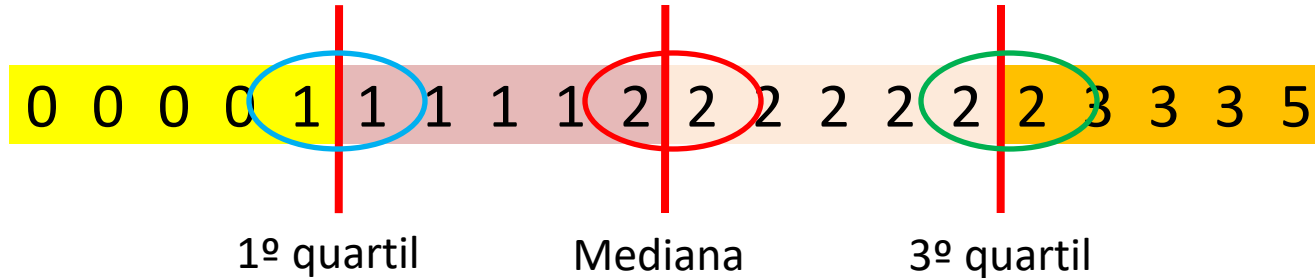
- Os quantis mais utilizados são:
- 1º Quartil = $q(0,25)$ = 25º Percentil;
- Mediana = $q(0,50)$ = 50º Percentil;
- 3º Quartil = $q(0,75)$ = 75º Percentil.

1º Quartil = 1

Mediana = 2

3º Quartil = 2

Exemplo - Par



N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.2	0.2	20%
1	5	0.25	0.45	25%
2	7	0.35	0.8	35%
3	3	0.15	0.95	15%
4	0	0	0.95	0%
5	1	0.05	1	5%
Total	20	1		100%

1º Quartil = 1

Mediana = 2

3º Quartil = 2

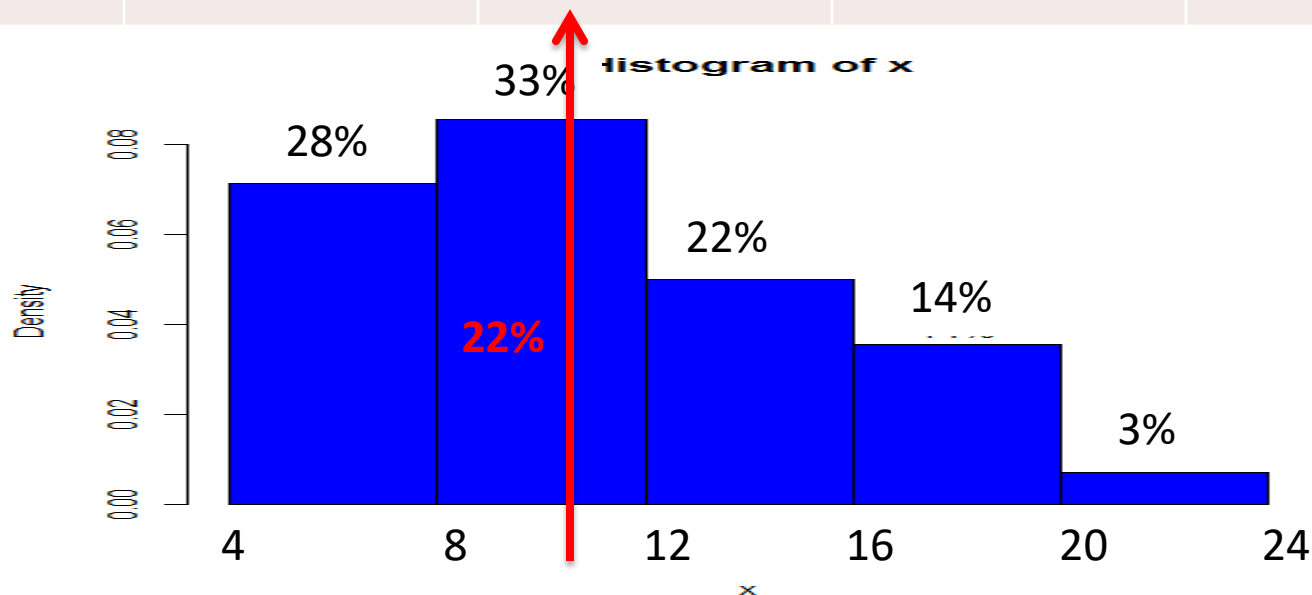
Exemplo - Ímpar

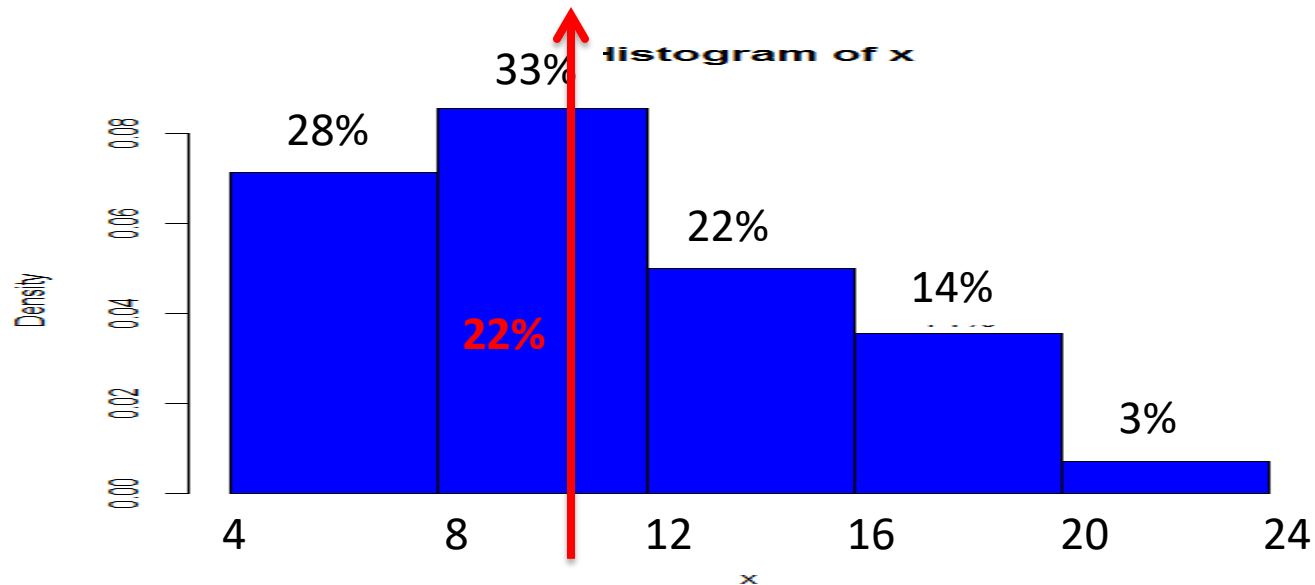
0 0 0 0 1 1 1 1 1 2 2 2 2 2 2 3 3 3

1º quartil Mediana 3º quartil

N. Filhos	Freq. Abs.	Freq. rela.	Freq. Acum.	%
0	4	0.21	0.21	21%
1	5	0.26	0.47	26%
2	7	0.37	0.84	37%
3	3	0.16	1.00	16%
Total	19	1		100%

Classe de salários	Freq	Freq relativa	Freq acum	Porcentagem
[4,00; 8,00)	10	$10/36 = 0,278$	0,278	27,78%
[8,00; 12,00)	12	$12/36 = 0,333$	0,611	33,33%
[12,00; 16,00)	8	$8/36 = 0,222$	0,833	22,22%
[16,00; 20,00)	5	$5/36 = 0,139$	0,972	13,89%
[20,00; 24,00)	1	$1/36 = 0,029$	1,000	2,78%
Total	36	1		100%

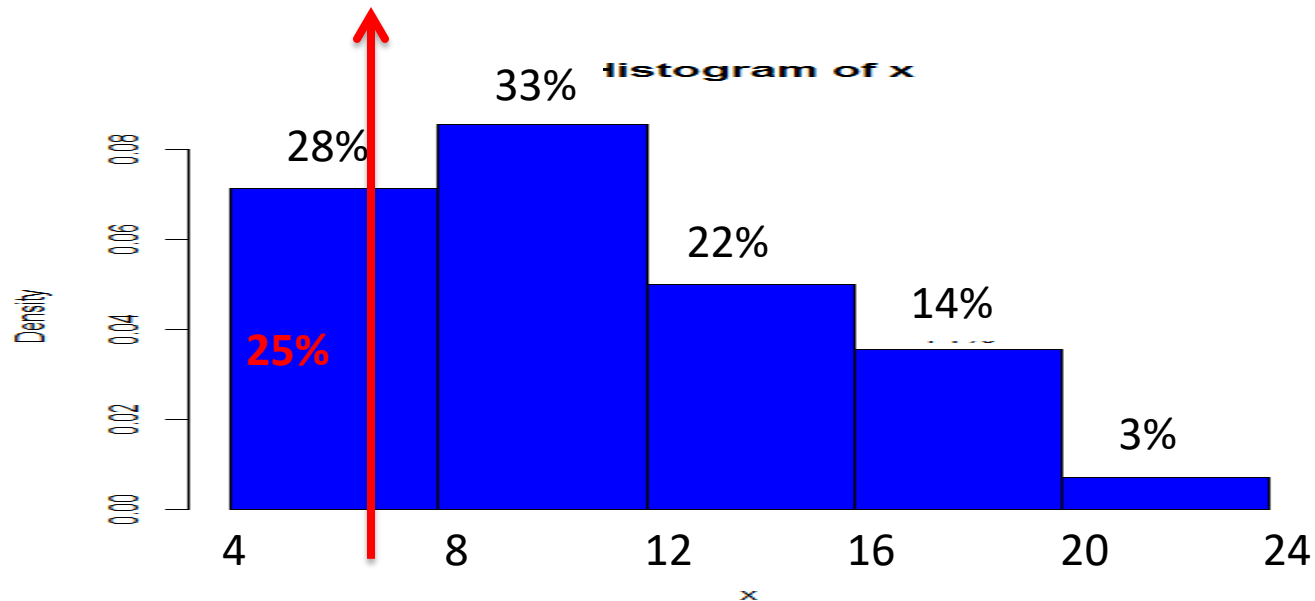




$$\frac{12,00 - 8,00}{33\%} = \frac{md - 8,00}{22\%}$$

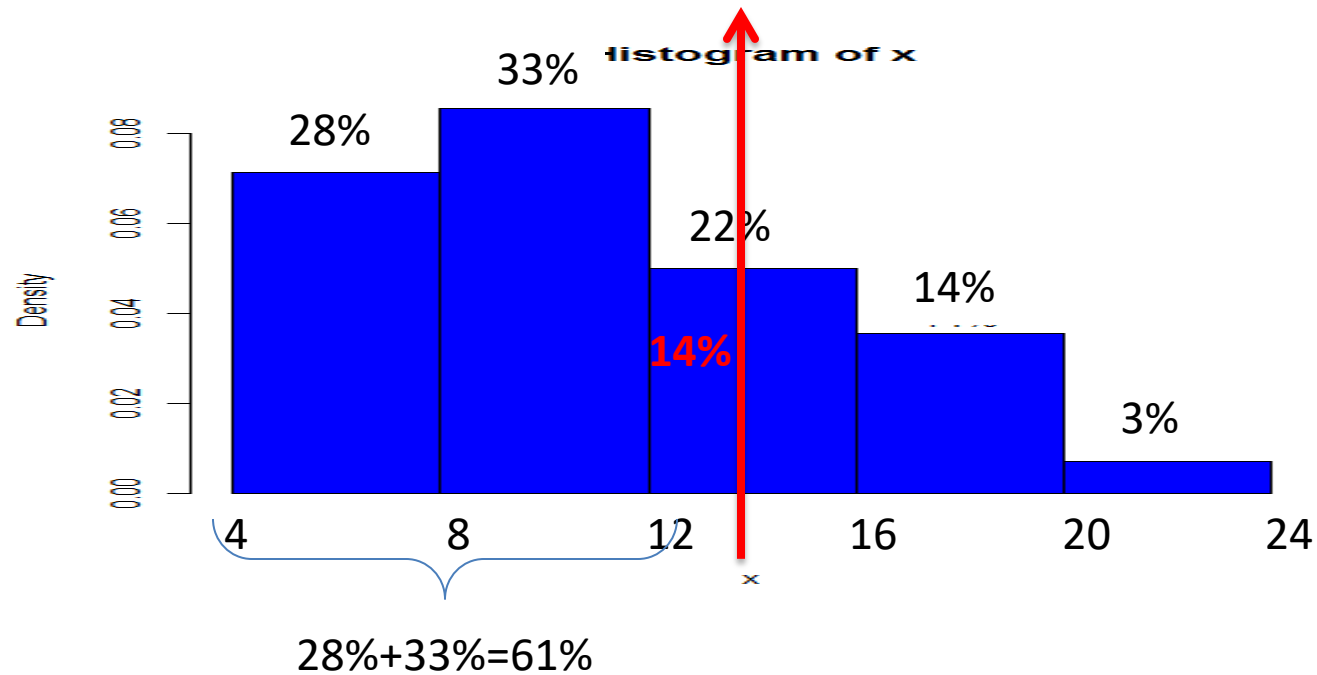
$$md - 8,00 = \frac{22\%}{33\%} 4$$

$$md = 8,00 + 2,67 = 10,67$$



$$\frac{q(0,25) - 4,00}{25\%} = \frac{8,00 - 4,00}{28\%}$$

$$q(0,25) = 4,00 + \frac{25}{28} 4,00 = 7,57$$



$$\frac{q(0,75) - 12,00}{14\%} = \frac{16,00 - 12,00}{22\%}$$

$$q(0,25) = 14,55$$

No Programa R

```
quantile(variavel,0.25)
```

```
quantile(variavel,0.5)
```

```
quantile(variavel,0.75)
```

```
quantile(variavel)
```

Medida de dispersão alternativa

- Uma medida de dispersão alternativa ao desvio padrão é a **distância interquartil**, consiste na diferença entre o terceiro e o primeiro quartil:

$$d_q = q(0,75) - q(0,25)$$

Resistência

- Os quartis são medidas de posição resistentes.
- Uma medida de posição ou dispersão é **resistente** quando for pouco afetada por mudanças de uma pequena porção de dados.
- A mediana é uma medida resistente, a média e o desvio padrão não são medidas resistentes.

Exemplo

- Considere as populações dos 20 municípios mais populosos de Minas Gerais, segundo o censo do IBGE de 2010.

Município	População		Município	População
Belo Horizonte	2.375.444		Sete Lagoas	214.071
Contagem	603.048		Divinópolis	213.076
Uberlândia	600.285		Santa Luzia	203.184
Juiz de Fora	517.872		Ibirité	159.026
Betim	377.547		Poços de Caldas	152.496
Montes Claros	361.971		Patos de Minas	138.836
Ribeirão das Neves	296.376		Teófilo Otoni	134.733
Uberaba	296.000		Pouso Alegre	130.586
Governador Valadares	263.594		Barbacena	126.325
Ipatinga	239.177		Sabará	126.219

- Medidas resumos:

Município	População		Município	População
Belo Horizonte	2.375.444		Sete Lagoas	214.071
Contagem	603.048		Divinópolis	213.076
Uberlândia	600.285		Santa Luzia	203.184
Juiz de Fora	517.872		Ibirité	159.026
Betim	377.547		Poços de Caldas	152.496
Montes Claros	361.971		Patos de Minas	138.836
Ribeirão das Neves	296.376		Teófilo Otoni	134.733
Uberaba	296.000		Pouso Alegre	130.586
Governador Valadares	263.594		Barbacena	126.325
Ipatinga	239.177		Sabará	126.219

Média= 376.493

Desvio padrão=481.825

3º quartil= 365.865

Mediana = 226.624

1º quartil= 149.081

Dist. interq.= 216.784

Sem BH

Média= 271.285

Desvio padrão=151.660

3º quartil= 329.174

Mediana = 214.071

1º quartil= 145.666

Dist. interq.= 183.508

Idéia de simetria

- Os cinco valores são importantes para se ter uma boa idéia da assimetria da distribuição dos dados:

$x_{(1)}$, q_1 , Mediana, q_3 , $x_{(n)}$

- Para uma distribuição simétrica deveríamos ter:

Dispersão inferior

Dispersão superior

1) $\text{Mediana} - x_{(1)} \approx x_{(n)} - \text{Mediana}$

2) $\text{Mediana} - q_1 \approx q_3 - \text{Mediana}$

3) $q_1 - x_{(1)} \approx x_{(n)} - q_3$

Distribuição assimétrica

1. $100.405 \neq 2.148.820$
2. $77.543 \neq 139.241$
3. $22.862 \neq 2.009.579$

Média= 376.493
Desvio padrão=481.825
3º quartil= 365.865
Mediana = 226.624
1º quartil= 149.081
Dist. interq.= 216.784
Min= 126.219
Max= 2.375.444

Média= 271.285
Desvio padrão=151.660
3º quartil= 329.174
Mediana = 214.071
1º quartil= 145.666
Dist. interq.= 183.508
Min= 126.219
Max= 603.048

Box Plot

- O Box Plot é o gráfico que contém os valores da mediana, 1º e 3º quartis, limite superior e inferior e observações discrepantes.

- O limite inferior é obtido por:

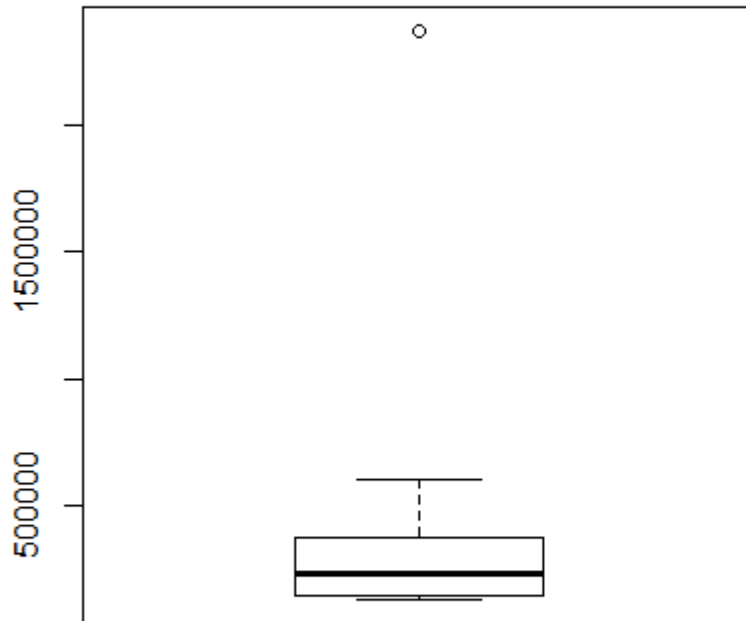
$$L_i = q_1 - (1,5)d_q$$

- O limite superior é obtido por:

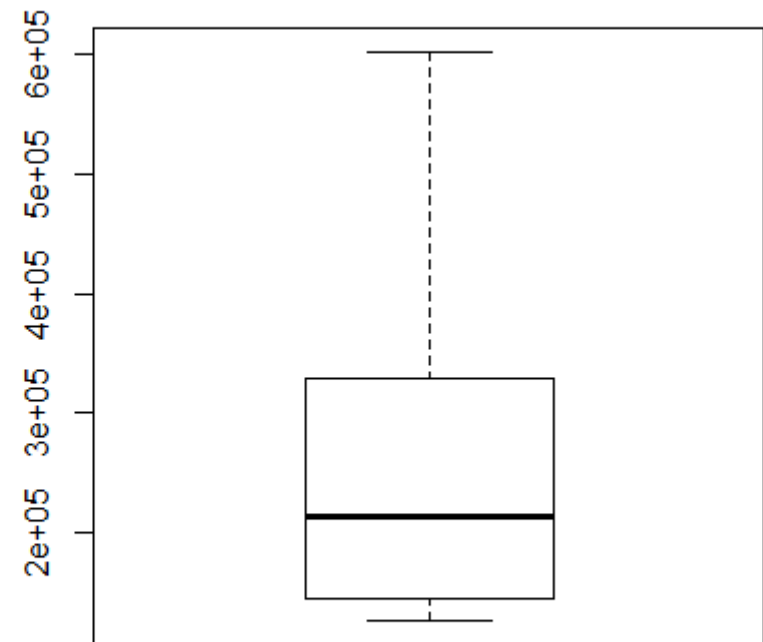
$$L_s = q_3 + (1,5)d_q$$

Distribuição assimétrica

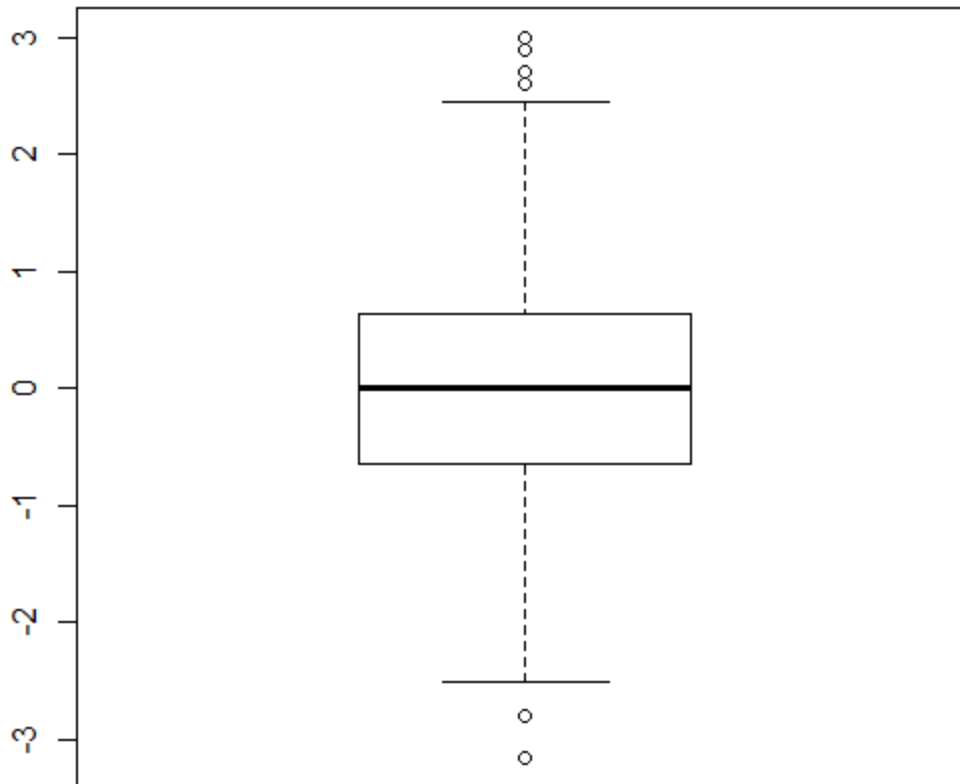
Com BH



Sem BH



Distribuição simétrica



No Programa R

IQR(variavel)

boxplot(variavel)

Exercício 1

- Obtenha os resumos descritivos que achar mais pertinentes para a variável faturamento.
- Use o programa R

Exercício 2

- O número de divórcios na cidade, de acordo com a duração do casamento, está representado na tabela abaixo:

Anos de casamento	No de divórcios
[0,6)	2800
[6,12)	1400
[12,18)	600
[18,24)	150
[24,30)	50

- 1) Qual a duração média dos casamentos? E a mediana?
- 2) Encontre a variância e o desvio padrão da duração dos casamentos.
- 3) Construa o histograma da distribuição.
- 4) Encontre o 1º e 3º quartil.
- 5) Qual o intervalo interquartil?
- 6) Construa o box-plot da distribuição.

Exercício 3

- O que acontece com a mediana, média e o desvio padrão de uma série de dados quando:
 1. Cada observação é multiplicada por 2?
 2. Soma-se 10 a cada observação?
 3. Subtrai-se a média geral de cada observação?
 4. Subtrai-se a média geral de cada observação e divide-se pelo desvio padrão?