



FEAUSP



## PARTE 1

Prof<sup>a</sup>. Dr<sup>a</sup>. Alessandra de Ávila Montini

## Programa de Pós-graduação do Departamento de Administração – PPGA



### Agenda



1

14:00  
15:30

Fundamentação Teórica

2

15:30  
15:45

*Coffee break*

3

15:45  
17:00

Fundamentação Teórica

4

17:00  
18:00

Exercícios de Fixação



**Os exercícios de fixação da sala de aula só poderão ser feitos na sala de aula no dia proposto.**



FEAUSP

## Critérios de Avaliação

## Frequência

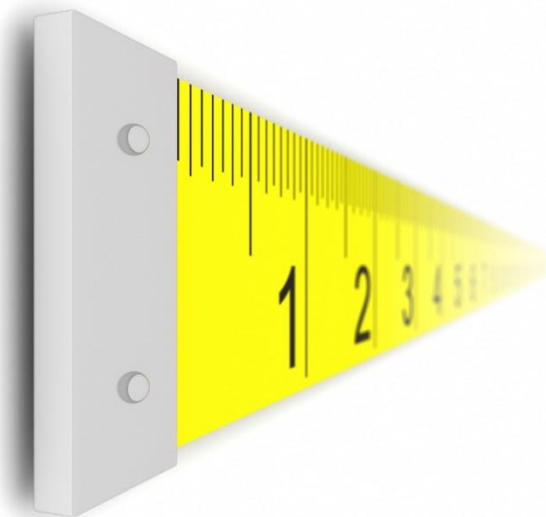
15 dias letivos

**3 faltas permitidas**

Deve-se ter 75 % de presença

## Atrasos

Hora limite para obtenção de presença na aula : 14 h 30 min



A → [8,50; 10,0]

B → [7,01; 8,49]

C → [5,00; 7,00]

R → [0,00; 4,99]

EXERCÍCIOS → 0,30

PROVA1 → 0,30

PROVA 2 → 0,40

TOTAL → 1,00

## Conceito

## Pesos



FEAUSP

## Programa



# Programa

- Análise Exploratória de Dados: Medidas de Posição e Dispersão
- Introdução à Probabilidade
- Distribuições Discretas de Probabilidade
- Distribuições Contínuas de Probabilidade
- Estimação por Ponto e por Intervalo
- Testes de Hipóteses
- Comparações Envolvendo Médias
- Comparações Envolvendo Proporções
- Análise de Variância
- Teste Qui-Quadrado
- Amostragem Aleatória Simples
- Amostragem Aleatória Sistemática
- Amostra Aleatória Estratificada





FEAUSP

## Referência Bibliográfica

Anderson, D. R., Sweeney, D. J. e Williams, T. A.  
Estatística Aplicada à Administração e Economia.  
Cengage Learning. Tradução da 6ª edição norte-  
americana. 3ª edição brasileira. 2014

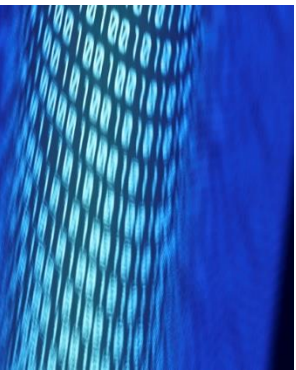
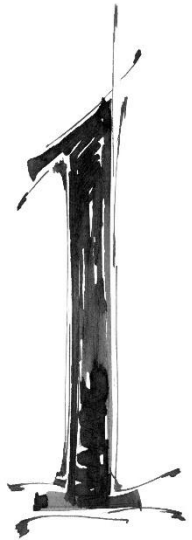




**FEAUSP**

Cronograma Regular





# ANÁLISE EXPLORATÓRIA DE DADOS



# Conteúdo

- Elaboração de Bancos de Dados;
- Tipos de Variáveis
  - Dados Qualitativos
  - Dados Quantitativos
- Distribuição de frequência
- Simetria e Assimetria
- Histograma
- População e Amostra
- Medidas de posição
  - Moda
  - Média Aritmética
  - Média Ponderada
  - Mediana
  - Percentil
  - Quartil
- Valor Discrepante
- Box-plot e Identificação de Outlier
- Mínimo e Máximo



# Conteúdo

- Medidas de variabilidade
  - Desvio
  - Desvio Médio Absoluto
  - Variância Populacional e Amostral
  - Desvio Padrão Populacional e Amostral
  - Coeficiente de Variação
- Cálculo da média e do desvio padrão amostral via HP12C
- Análise Bivariada
- Gráfico de Dispersão
- Coeficiente de Correlação Linear
- Covariância

# Banco de Dados

# Banco de Dados - EXCEL

Quando abre-se um banco de dados em softwares estatísticos, geralmente, espera-se que o banco de dados seja formado como o apresentado. Na primeira coluna inicia-se a base de dados e na primeira linha adiciona-se o nome da variável.

M5											
	A	B	C	D	E	F	G	H	I	J	K
1	Código do Cliente	Sexo	Estado Civil	Estado de Residência	Possui Cartão de Crédito	Idade	Rendimento Total	Salário	Limite de Crédito Imediato	Valor Total do Patrimônio	Limite do Cheque Especial
2	1	F	viúvo	RJ	sim	81	6800	6800	380	299109	2000
3	2	F	viúvo	RJ	sim	35	5000	5000	1000	120000	1000
4	3	F	viúvo	RJ	sim	39	6320	6320	1550	100000	1640
5	4	F	divorciado	RJ	não	70	10736	5214	400	100000	500
6	5	F	casado	SP	não	54	6000	6000	1790	171745	3600
7	6	M	solteiro	SP	sim	64	15000	15000	3000	561138	10000
8	7	M	casado	SP	não	69	37000	22000	1000	2593588	4000
9	8	F	casado	SP	não	68	10527	4027	3000	350000	5000
10	9	M	casado	SP	não	30	8000	8000	3000	200000	3350
11	10	M	casado	RJ	não	72	7825	7825	3000	120000	3000
12	11	F	casado	SP	não	73	7890	7000	3000	17939	5000
13	12	F	divorciado	RJ	não	72	4200	4200	3000	507000	4000

<http://www.denatran.gov.br/motazu14.htm>

Fonte: Ministério das Cidades, Departamento Nacional de Trânsito - DENATRAN, Sistema Nacional de Registro de Veículos - RENAVAM.

Profa. Dra. Alessandra de Ávila Montini



# Tipo de Variável

Após a criação de um banco de dados é muito importante identificar o tipo de variável pois nem todas as medidas descritivas podem ser calculadas para todo tipo de variável.

O banco de dados de dados apresentado relacionado às informações de cadastro de clientes possui algumas variáveis qualitativas (como as variáveis: sexo, estado civil, estado de residência e possui cartão de crédito) e algumas variáveis quantitativas (rendimento total, salário, limite de crédito imediato, valor total do patrimônio e limite do cheque especial).

M5 <span>fx</span>											
	A	B	C	D	E	F	G	H	I	J	K
1	Código do Cliente	Sexo	Estado Civil	Estado de Residência	Possui Cartão de Crédito	Idade	Rendimento Total	Salário	Limite de Crédito Imediato	Valor Total do Patrimônio	Limite do Cheque Especial
2	1	F	viúvo	RJ	sim	81	6800	6800	380	299109	2000
3	2	F	viúvo	RJ	sim	35	5000	5000	1000	120000	1000
4	3	F	viúvo	RJ	sim	39	6320	6320	1550	100000	1640
5	4	F	divorciado	RJ	não	70	10736	5214	400	100000	500
6	5	F	casado	SP	não	54	6000	6000	1790	171745	3600
7	6	M	solteiro	SP	sim	64	15000	15000	3000	561138	10000
8	7	M	casado	SP	não	69	37000	22000	1000	2593588	4000
9	8	F	casado	SP	não	68	10527	4027	3000	350000	5000
10	9	M	casado	SP	não	30	8000	8000	3000	200000	3350
11	10	M	casado	RJ	não	72	7825	7825	3000	120000	3000
12	11	F	casado	SP	não	73	7890	7000	3000	17939	5000
13	12	F	divorciado	RJ	não	70	10736	5214	400	100000	500

# Tipo de Variável

As variáveis de um banco de dados podem se classificadas como qualitativas ou quantitativas.

Não pode ser utilizada em cálculos, pois representam uma qualidade

- Qualitativa
  - Ordinal
  - Nominal
- Quantitativa
  - Discreta
  - Contínua

# Tipo de Variável

Exemplo de variável qualitativa ordinal.

## Exemplos:

- Nível de escolaridade
- Tamanho (pequeno, médio , grande)
- Nível de risco (alto risco, médio risco, baixo risco)



# Tipo de Variável

Exemplo de variável qualitativa nominal.

## Exemplos:

- Sexo
- Estados brasileiros
- Raça
- Setor de atuação (indústria, comércio, serviços)

# Tipo de Variável

Exemplo de variável quantitativa discreta.

## Exemplos:

- Número de filhos
- Número de dias para pagamento
- Número de clientes
- Número de unidades vendidas

# Tipo de Variável

Exemplo de variável quantitativa contínua.

## Exemplos:

- Salário
- Faturamento
- Cotação do dólar

# Distribuição de Frequência



# Distribuição de Frequência

A tabela apresenta a **frequência absoluta** e a **frequência relativa** relacionada aos cursos de aperfeiçoamento realizados por 500 executivos de uma empresa.

Cada linha da tabela é considerada uma **categoria**.

A **frequência absoluta** é o número de observações na categoria.

A **frequência relativa** é obtida por meio da divisão da **frequência absoluta** pelo **total de observações**.

Cursos de Aperfeiçoamento	Frequência Absoluta	Frequência Relativa
1	100	0,20
2	120	0,24
3	50	0,10
4	80	0,16
5	100	0,20
6	50	0,10
<b>Total</b>	<b>500</b>	

A soma das frequências relativas deve ser igual a um.

# Distribuição de Frequência

A **frequência relativa percentual** é obtida por meio da multiplicação da **frequência relativa** por 100.

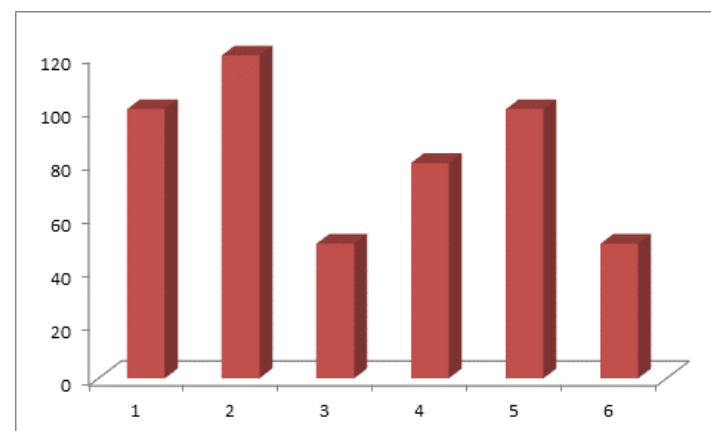
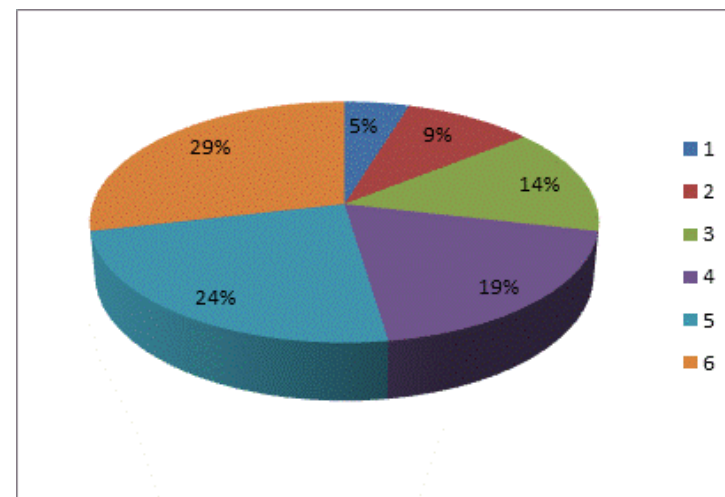
<b>Cursos de Aperfeiçoamento</b>	<b>Frequência Absoluta</b>	<b>Frequência Relativa</b>	<b>Frequência Relativa Percentual (%)</b>
1	100	0,20	20,0%
2	120	0,24	24,0%
3	50	0,10	10,0%
4	80	0,16	16,0%
5	100	0,20	20,0%
6	50	0,10	10,0%
<b>Total</b>	<b>500</b>		

# Distribuição de Frequência

A **frequência relativa percentual** pode ser representada por meio de um gráfico de Pizza ou por gráfico de Barra.

<b>Cursos de Aperfeiçoamento</b>	<b>Frequência Relativa Percentual (%)</b>
1	20,00%
2	24,00%
3	10,00%
4	16,00%
5	20,00%
6	10,00%
<b>Total</b>	

Histograma - quando a variável é contínua.



# Distribuição de Frequência

As variáveis quantitativas discretas sempre podem ser apresentadas por meio de uma distribuição de frequência.

As variáveis qualitativas sempre podem ser apresentadas por meio de uma distribuição de frequência.



# Distribuição Simétrica e Assimétrica

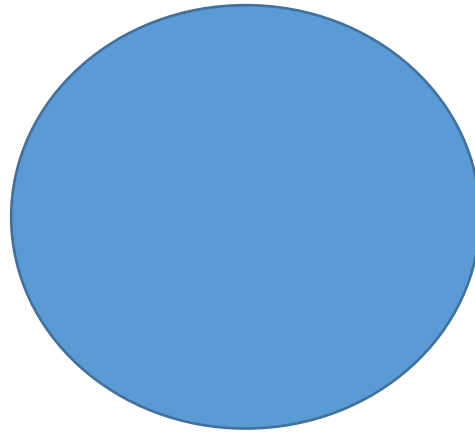
# Simetria

O conceito de simetria é intuitivo.

Nota-se que a borboleta é simétrica



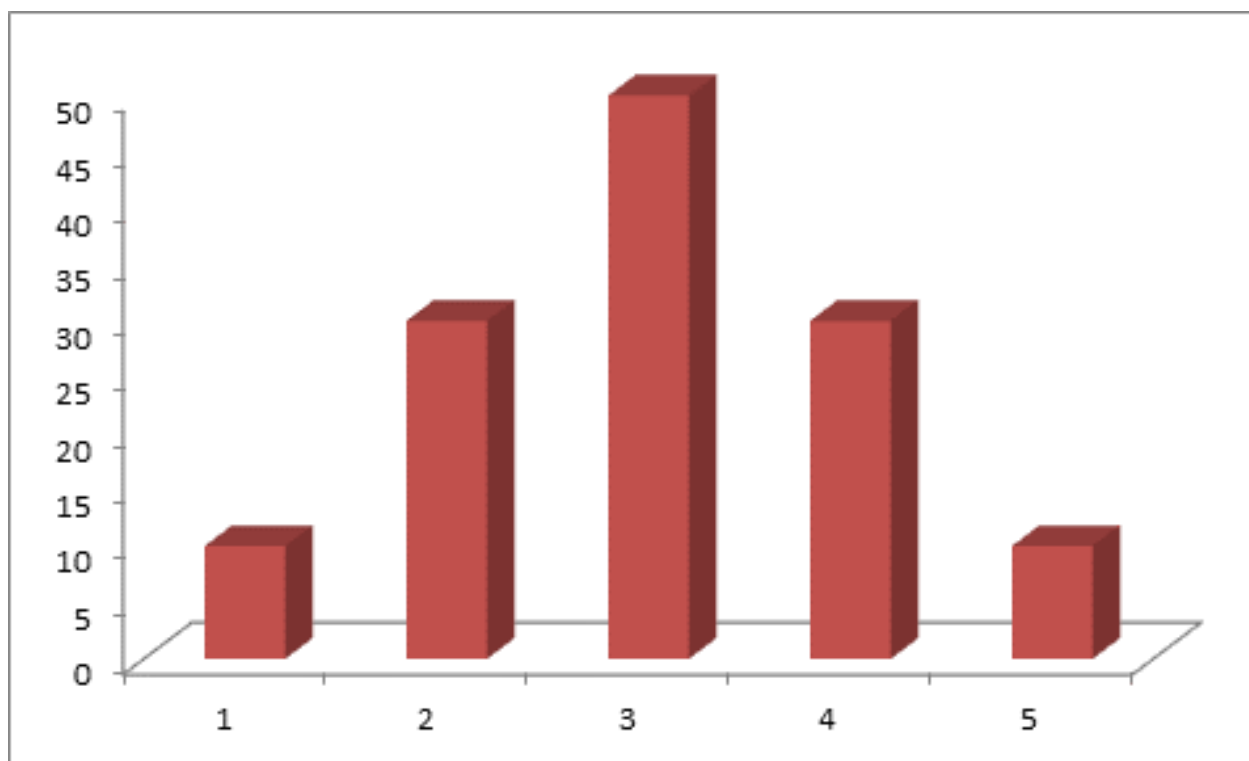
O círculo é simétrico.



# Distribuição Simétrica

O gráfico de barra apresenta a **frequência absoluta** (número de pedidos realizados) por cinco vendedores de uma determinada empresa.

Neste exemplo, pode-se dizer que a **distribuição dos dados é simétrica**.

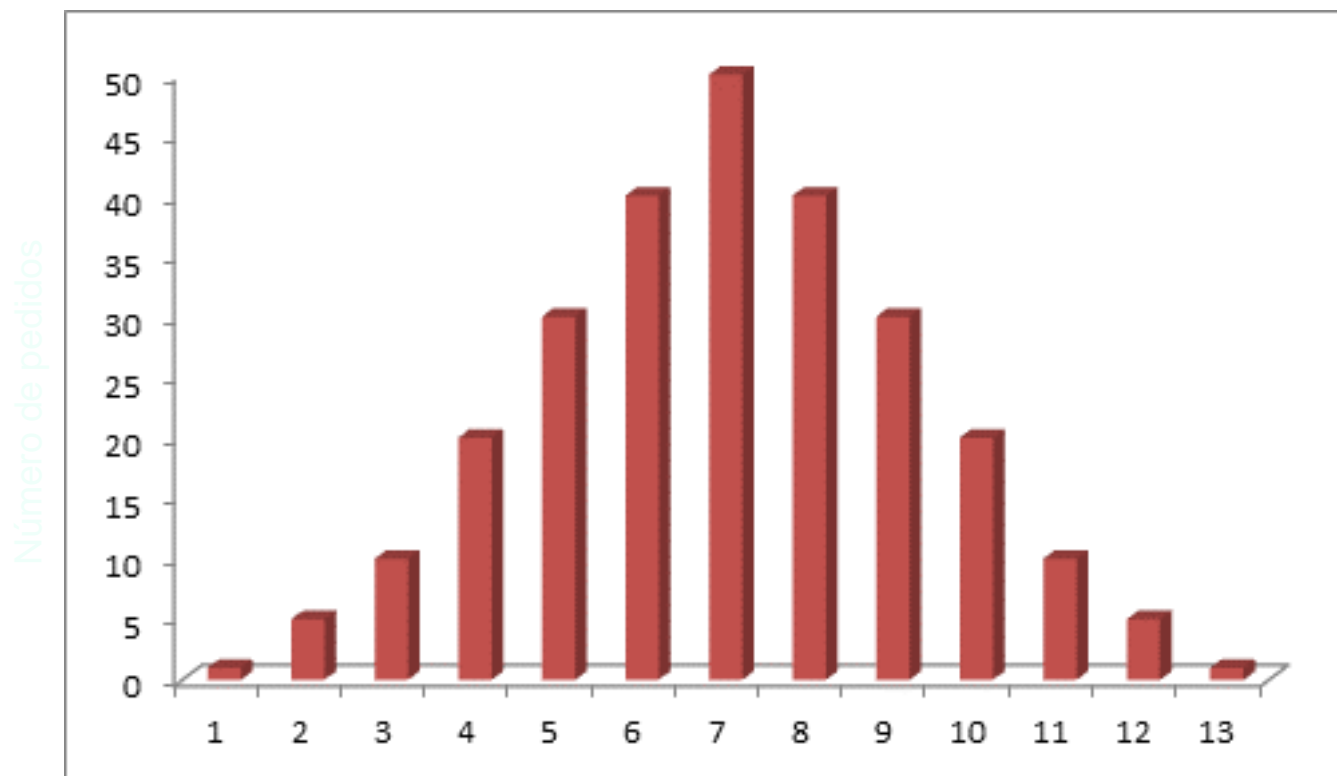




# Distribuição Simétrica

O gráfico de barra apresenta a **frequência absoluta** (número de pedidos realizados) por treze vendedores de uma determinada empresa.

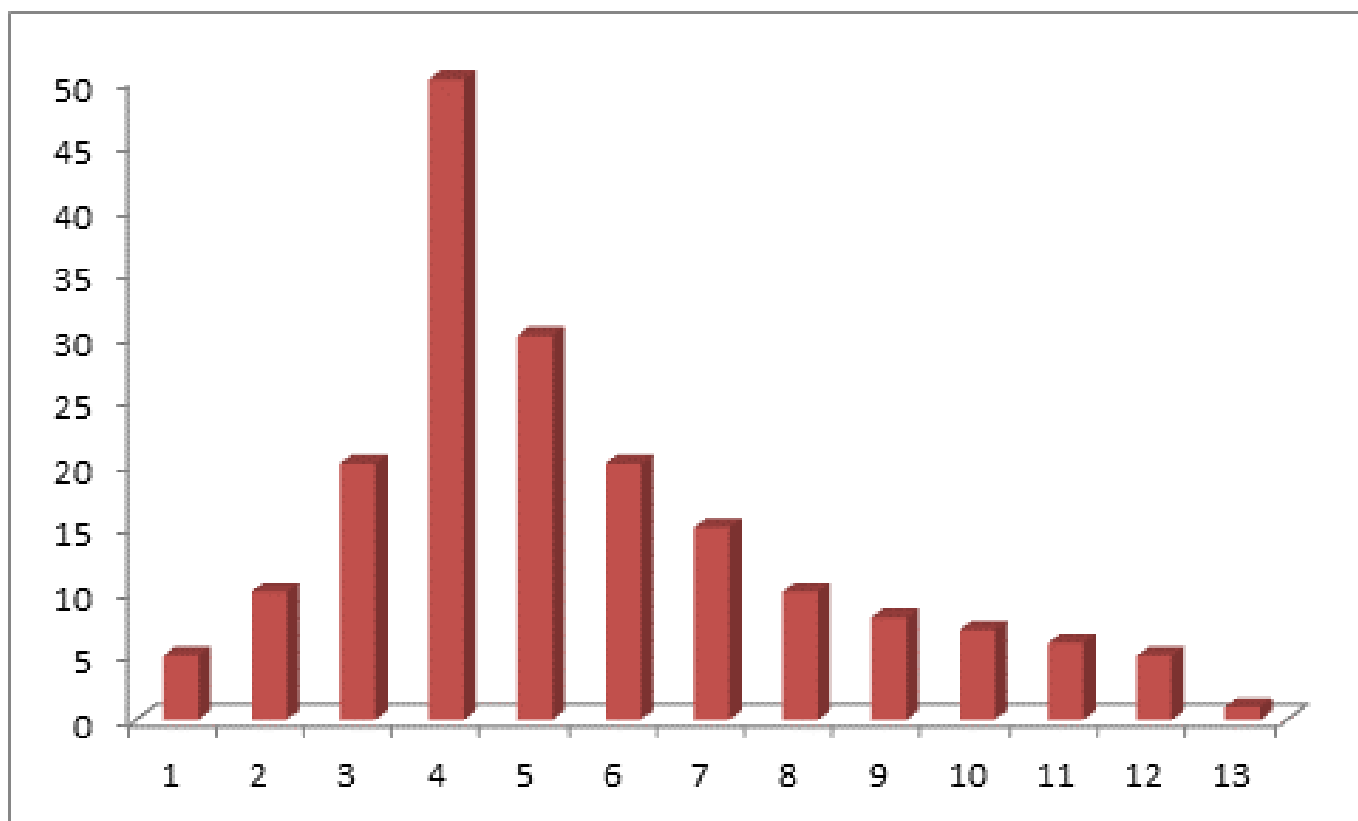
Neste exemplo, pode-se dizer que a **distribuição dos dados é simétrica**.



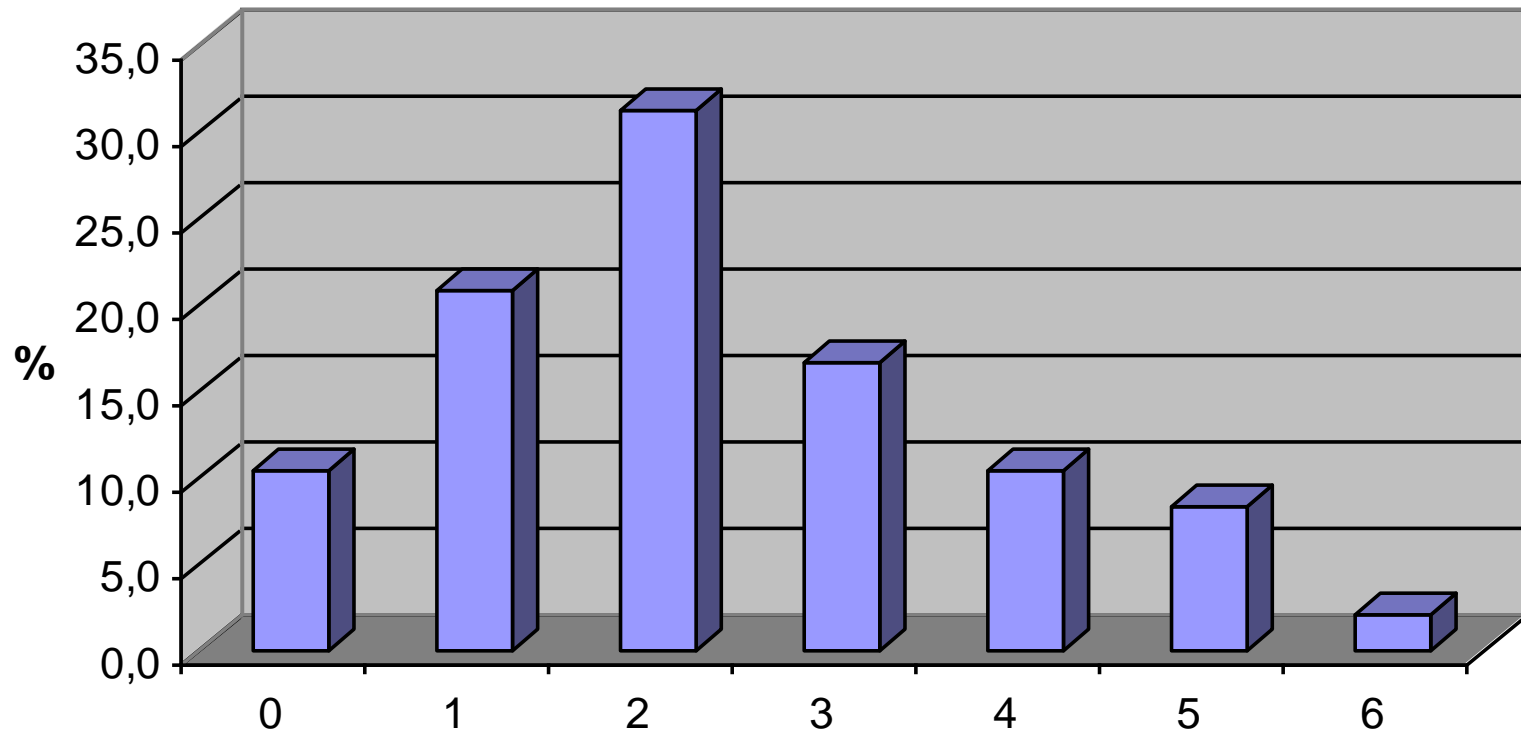
# Distribuição Assimétrica à Direita

O gráfico de barra apresenta a **frequência absoluta** (número de carros de luxo vendidos) por treze concessionárias.

Neste exemplo, pode-se dizer que a **distribuição dos dados é assimétrica à direita**.



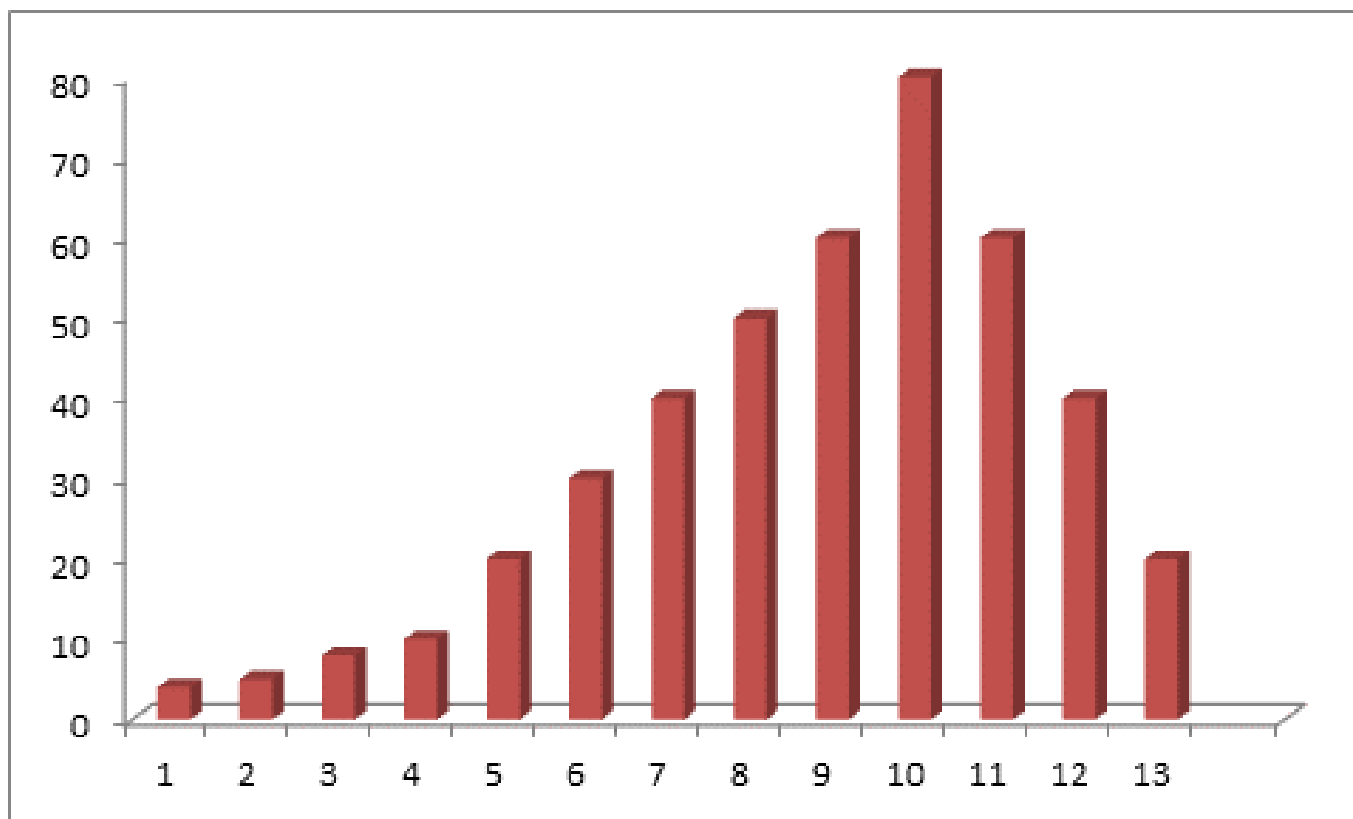
# Distribuição Assimétrica à Direita



# Distribuição Assimétrica à Esquerda

O gráfico de barra apresenta a **frequência absoluta**, ou seja, o número de sinistros ocorridos nos últimos 10 anos para a frota de 13 empresas.

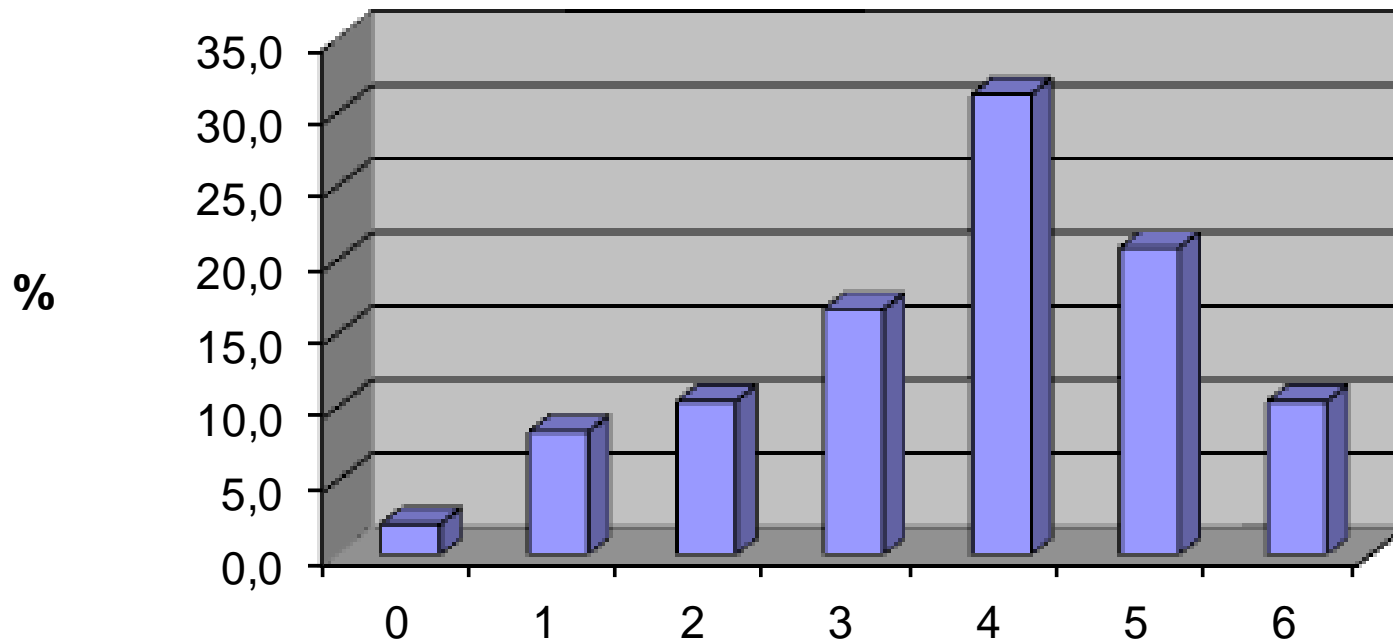
Neste exemplo, pode-se dizer que a **distribuição dos dados é assimétrica à esquerda**.



# Distribuição Assimétrica à Esquerda

O gráfico de barra apresenta a **frequência absoluta, ou seja**, o número de medalhas obtidos por 6 atletas nos últimos 5 anos.

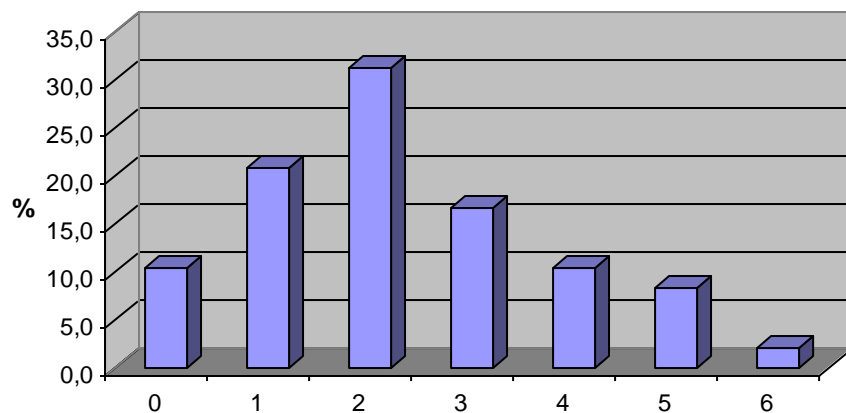
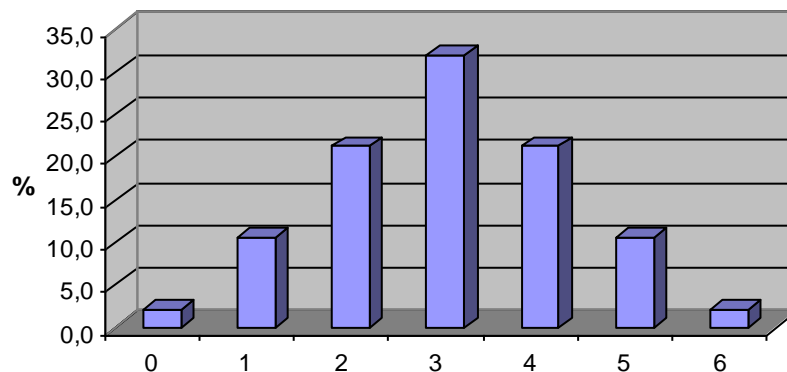
Neste exemplo, pode-se dizer que a **distribuição dos dados é assimétrica à esquerda**.





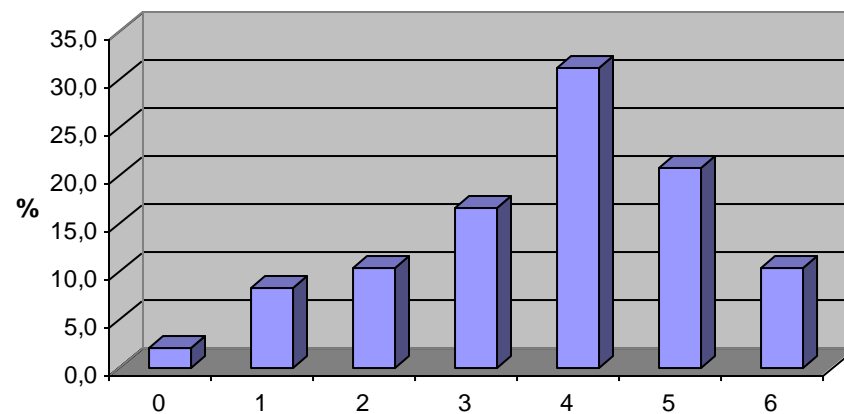
## Distribuição Simétrica

$A=0$



$A > 0$  – Distribuição inclinada para a direita

Distribuição Assimétrica à Direita



$A < 0$  - Distribuição inclinada para a esquerda

Distribuição Assimétrica à Esquerda

# Histograma

# Histograma

A tabela apresenta o salário anual de 500 clientes de uma empresa.

Para obter o histograma de uma variável quantitativa pode-se obter a frequência absoluta, a frequência relativa e gerar o histograma com base na frequência relativa.

Cliente	Salário Anual
1	R\$ 51.814,00
2	R\$ 52.669,70
3	R\$ 51.780,30
4	R\$ 51.587,90
⋮	⋮
⋮	⋮
⋮	⋮
500	R\$ 51.752,00



**Base de Dados**

Salário anual	Frequência Absoluta	Frequência Relativa
49.500,00 a 49.999,99	2	0,004
50.000,00 a 50.499,99	16	0,032
50.500,00 a 50.999,99	52	0,104
51.000,00 a 51.499,99	101	0,202
51.500,00 a 51.999,99	133	0,266
52.000,00 a 52.499,99	110	0,220
52.500,00 a 52.999,99	54	0,108
53.000,00 a 53.499,99	26	0,052
53.500,00 a 53.999,99	6	0,012
Total	500	1

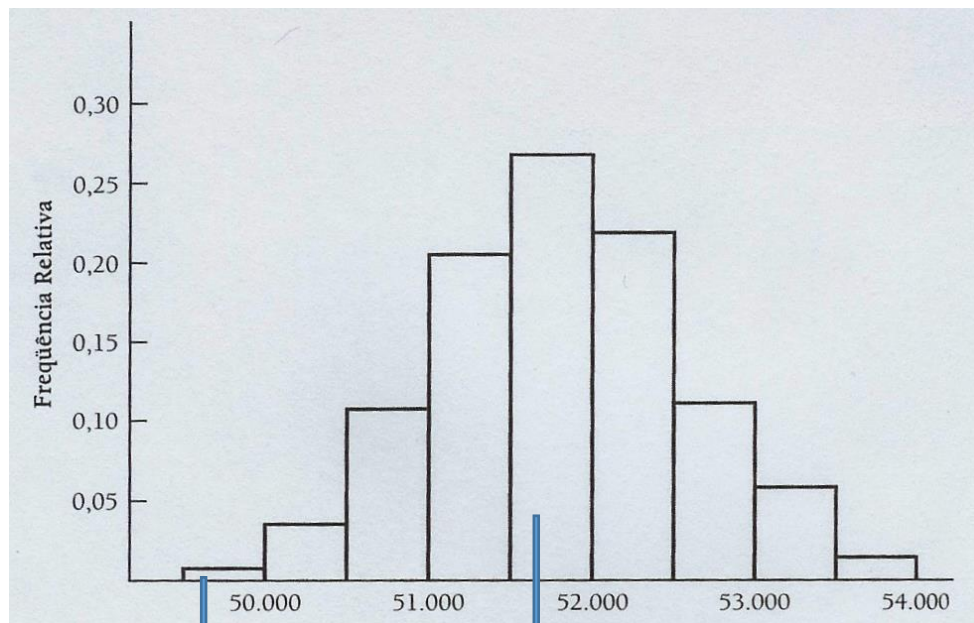


**Nota-se que 2 clientes da base de dados original possuem salário anual entre R\$ 49.500,00 e R\$ 49.999,99.**

# Histograma

As linhas da Tabela de Frequência são utilizadas para formar as barras do histograma e a altura da barra é obtida por meio da frequência relativa.

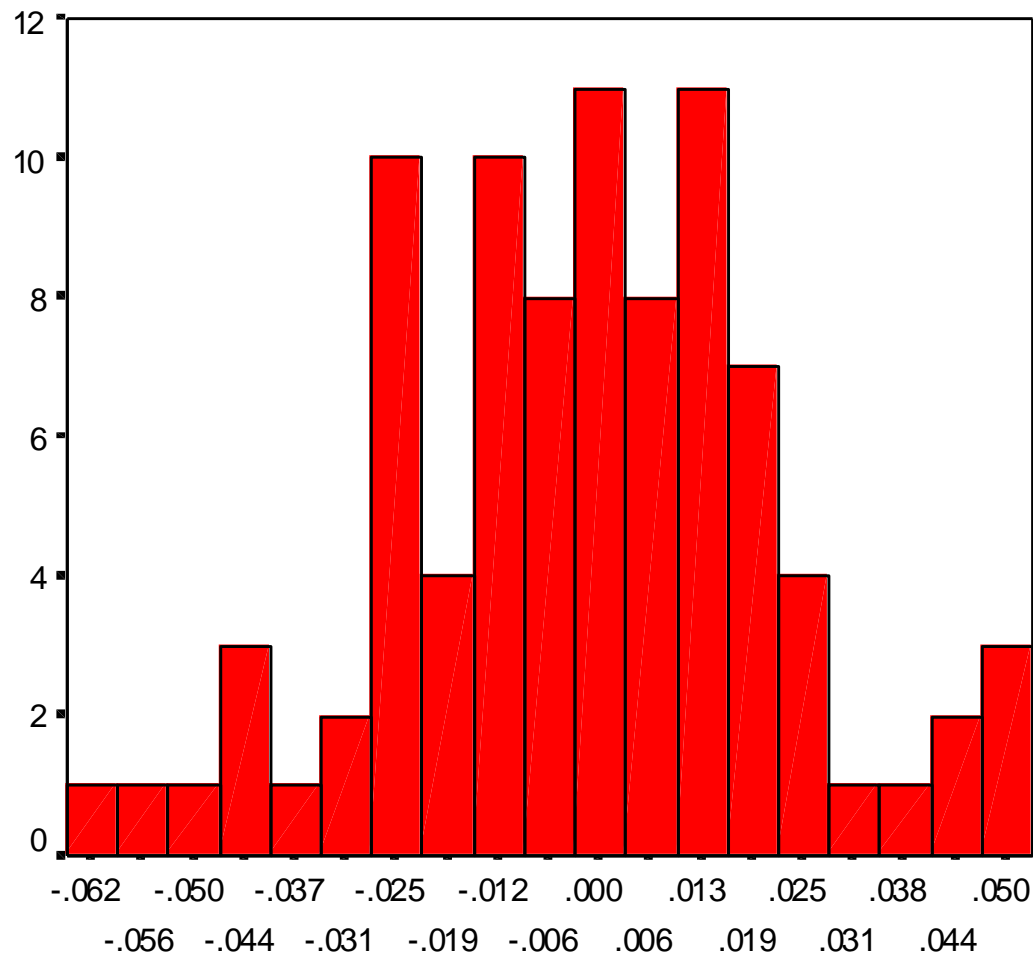
Salário anual	Frequência Absoluta	Frequência Relativa
49.500,00 a 49.999,99	2	0,004
50.000,00 a 50.499,99	16	0,032
50.500,00 a 50.999,99	52	0,104
51.000,00 a 51.499,99	101	0,202
51.500,00 a 51.999,99	133	0,266
52.000,00 a 52.499,99	110	0,220
52.500,00 a 52.999,99	54	0,108
53.000,00 a 53.499,99	26	0,052
53.500,00 a 53.999,99	6	0,012
Total	500	1



**Entre R\$ 49.500 e R\$ 49.999,99 há 0,004 ou 0,4 % dos clientes da base de dados original.**

**Entre R\$ 51.500 e R\$ 51.999,99 há 0,266 ou 26,6 % dos clientes da base de dados original.**

# Histograma



**Profa. Dra. Alessandra de Ávila Montini**



# População e Amostra

A **população** é formada por **todas as observações** do universo de referência.

O **tamanho** da população será denotado por **N**.

**Todos os aviões de um país**



**Todos os automóveis de uma cidade**



**Todos os peixes de um lago**



**Todos os brasileiros**



# Amostra

A amostra é formada por qualquer parte de uma população.

O tamanho da amostra será denotado por  $n$ .

**Alguns aviões do país**



**Alguns automóveis da cidade**



**Alguns peixes do lago**



**Alguns brasileiros**





# Análise Univariada

Na análise univariada o objetivo é fazer uma análise estatística para cada variável individualmente sem estudar a relação entre duas ou mais variáveis.

# Medidas de Posição e de Variabilidade



As medidas de posição e de variabilidade apresentadas só podem ser calculadas para variáveis quantitativas.

As medidas de posição e de variabilidade apresentadas **NÃO PODEM** ser calculadas para variáveis qualitativas.

## Medidas de Posição

- Moda
- Média Aritmética
- Média Ponderada
- Mediana
- Percentil
- Quartil
- Mínimo
- Máximo

## Medidas de Variabilidade

- Desvio
- Desvio Médio
- Variância Populacional
- Variância Amostral
- Desvio Padrão Populacional
- Desvio Padrão Amostral
- Coeficiente de Variação

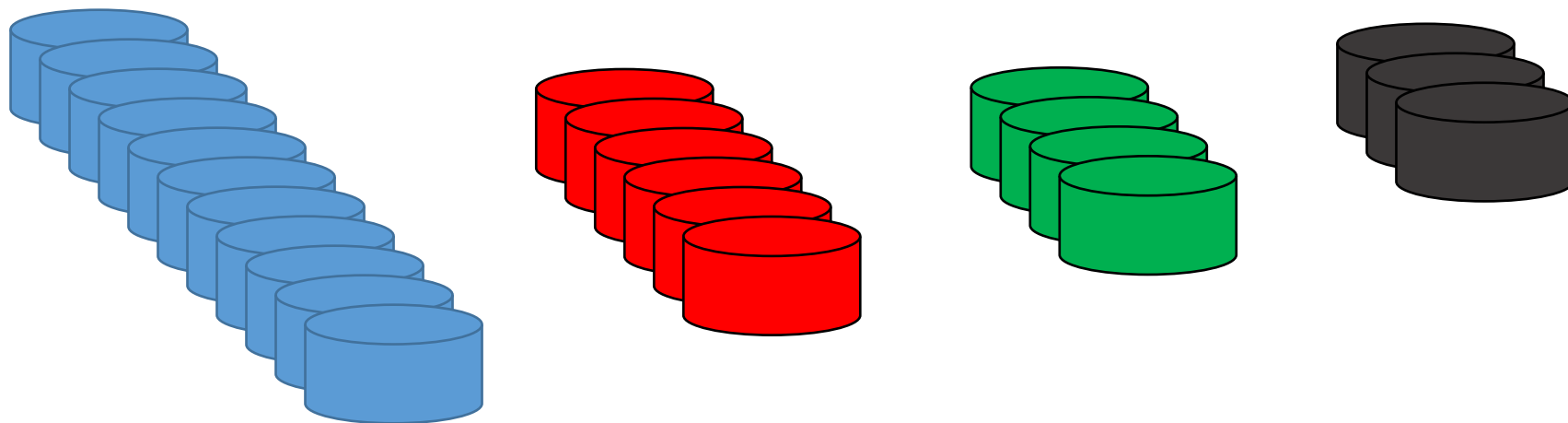
# Medidas de Posição

# Moda

# Moda

**Moda** é a **observação mais frequente** em um conjunto de observações.

Nesse conjunto de observações há **11** unidades na cor azul, **6** unidades na cor vermelha, **4** unidades na cor verde e **3** unidades na cor marrom. Desta forma a **moda é a cor azul** pois é a cor que aparece com a maior frequência.



# Moda

Um executivo deseja fazer uma análise no preço da concorrência com o objetivo de definir o preço a ser praticado em seu produto (TV LG 3D LED 47).



Por meio de uma busca no site Buscapé obteve-se o resultado apresentado. Nota-se que o preço mais praticado, ou seja, a moda é o valor R\$ 2.339,96



**E-bit Excelente** ★★★★★  
Avaliada por 10.735 pessoas  
[Mais detalhes da loja](#)

Preço:

**R\$ 2.339,96**

ou 10x R\$ 265,91  
com acréscimo



**E-bit Excelente** ★★★★★  
Avaliada por 236.286 pessoas  
[Mais detalhes da loja](#)

Preço:

**R\$ 2.184,05**



**E-bit Excelente** ★★★★★  
Avaliada por 14.090 pessoas  
[Mais detalhes da loja](#)

Preço:

**R\$ 2.339,96**



**E-bit Excelente** ★★★★★  
Avaliada por 21.945 pessoas  
[Mais detalhes da loja](#)

Preço:

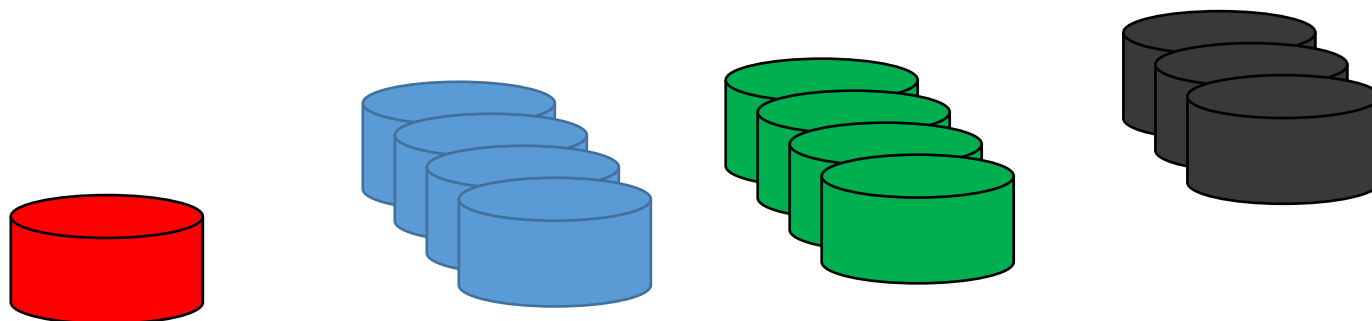
**R\$ 2.393,15**

# Moda

Uma base de dados pode ter duas modas. Neste caso tem-se uma distribuição **Bi-modal**.

Nesse conjunto de observações há **1** unidades na cor vermelha, **4** unidades na cor azul, **4** unidades na cor verde e **3** unidades na cor marrom.

Desta forma as **cores azul e verde** são consideradas **moda** pois são as cores que aparecem com a maior frequência.

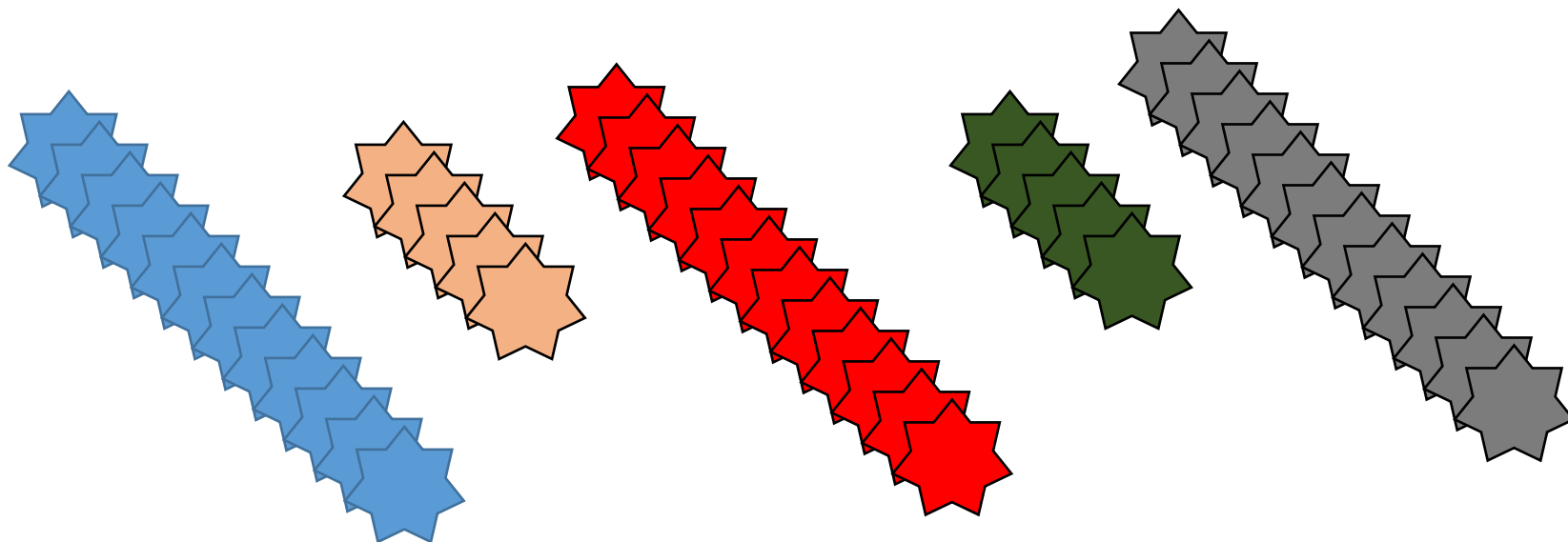




# Moda

Quando uma base de dados possui três ou mais observações que aparecem com frequência máxima tem-se uma distribuição **Multi-modal**.

Nesse conjunto de observações há **12** unidades na cor azul, **5** unidades na cor rosa, **12** unidades na cor vermelha, **5** unidades na cor marrom e **12** unidades na cor verde. Desta forma as **cores azul, vermelho e verde** são consideradas **moda** pois são as cores que aparecem com a maior frequência.



# Média

# Média

A palavra **média** pode ser utilizada em várias situações.

# Média





# Média

Na estatística a **média** é obtida a partir da soma das observações dividindo-se pelo total de observações.

# Média Aritmética

A **média** é a medida de posição mais utilizada.

A **média aritmética**, quando calculada para dados amostrais, é denotada por  $\bar{X}$

A **média aritmética**, quando calculada para dados populacionais, é denotada por  $\mu$

## Exemplo

Cálculo da **média** aritmética para o conjunto de dados:

$$30,3 + 31,0 + 31,3 + 31,5 + 31,8 + 31,3 + 31,3 + 31,1 + 31,2$$

A média é dada por:

$$\bar{X} = \frac{30,3 + 31,0 + 31,3 + 31,5 + 31,8 + 31,3 + 31,3 + 31,1 + 31,2}{9} = 31,2$$



# Média Aritmética

## Média do Enem 2013.

### **Veja as 20 escolas com as maiores médias nas PROVAS OBJETIVAS do Enem 2013**

- 1º) Colégio Objetivo Integrado (São Paulo/SP) - privada - média 741,94
- 2º) Colégio Bernoulli - unidade Lourdes (Belo Horizonte/MG) - privada - média 722,64
- 3º) Colégio e Curso Ponto de Ensino (Rio de Janeiro/RJ) - privada - média 720,02
- 4º) Colégio Vértice Unidade II (São Paulo/SP) - privada - média 715,41
- 5º) Colégio Santo Antônio (Belo Horizonte/MG) - privada - média 713,44
- 6º) Instituto Dom Barreto (Teresina/PI) - privada - média 713,39
- 7º) São Bento (Rio de Janeiro/RJ) - privada - média
- 8º) Colégio Ari de Sá - unidade Major Facundo (Fortaleza/CE) - privada - média 710,67
- 9º) Colégio Elite Vale do Aço (Ipatinga/MG) - privada - média 707,57
- 10º) Colegium (Belo Horizonte/MG) - privada - média 707,55
- 11º) SEB COC Unidade Álvares Cabral (Ribeirão Preto/SP) - privada - média 707,14
- 12º) Colégio de Aplicação da UFV - Coluni (Viçosa/MG) - federal - média 702,99
- 13º) Colégio e Curso Ponto de Ensino (Niterói/RJ) - privada - média 702,67
- 14º) Móbile Colégio (São Paulo/SP) - privada - média 702,18
- 15º) Colégio Santo Agostinho (Belo Horizonte/MG) - privada - média 701,67
- 16º) Colégio Olimpo (Brasília/DF) - privada - média 701,23
- 17º) Colégio Lerote Ltda (Teresina/PI) - privada - média 701,09
- 18º) Colégio Farias Brito - unidade central (Fortaleza/CE) - privada - média 696,35
- 19º) Colégio e Curso Ponto de Ensino (Rio de Janeiro/RJ) - privada - média 695,19
- 20º) Colégio Magnum Agostiniano - unid. Nova Floresta (Belo Horizonte/MG) - privada - 694,80

# Média Ponderada

A média ponderada deve ser calculada quando as observações possuem pesos diferentes.

Os pesos devem ter valores variando entre 0 e 1. A soma de todos os pesos deve ser igual a 1.

## Exemplo

Cálculo da média ponderada entre as notas de exercício e da prova.

$$\bar{X}_p = p_1 * \textit{exercício} + p_2 * \textit{prova}$$



Peso da nota de  
exercício



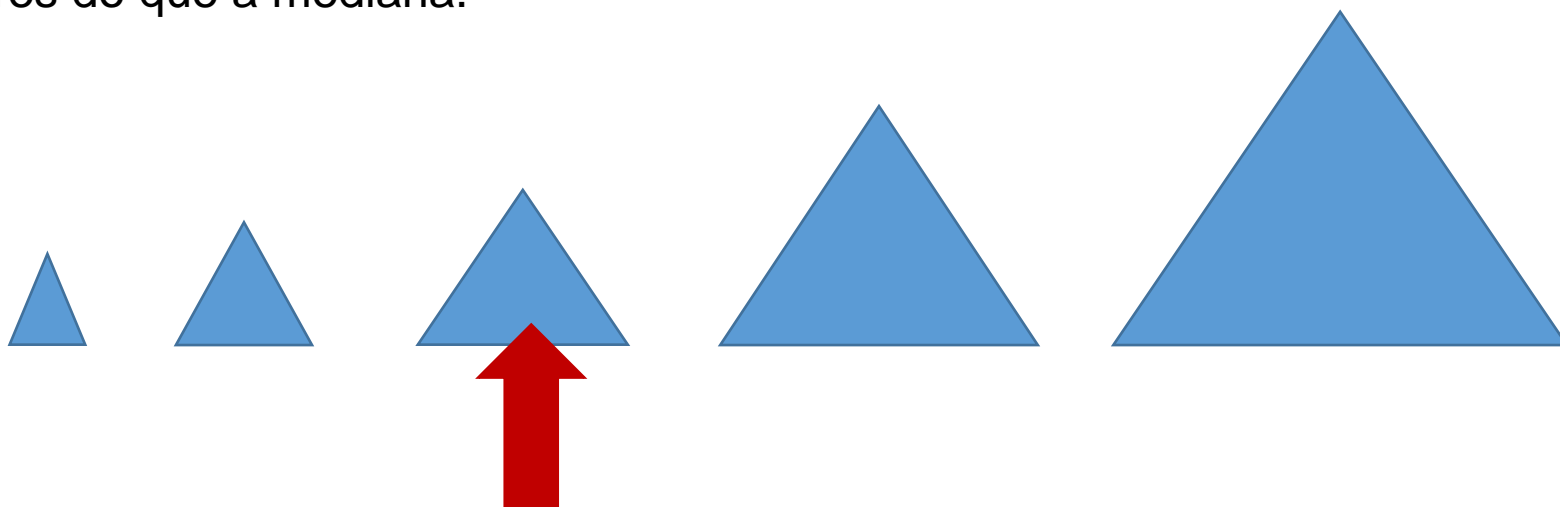
Peso da nota da prova

# Mediana

# Mediana

A **mediana** é a observação que ocupa a **posição central** de um conjunto de observações.

50 % das observações são maiores do que a mediana e 50 % das observações são menores do que a mediana.



Este triângulo representa a **mediana** pois 50 % dos triângulos são maiores do que ele e 50 % dos triângulos são menores do que ele.

# Mediana

Cálculo da mediana para a base de dados de salários.

## Salários

R\$ 23.000

R\$ 18.000

R\$ 25.000

R\$ 15.000

R\$ 12.000

### DICA

Para se obter a mediana inicialmente deve-se ordenar a base de dados em ordem crescente ou decrescente.

# Mediana

Cálculo da mediana para a base de dados de salários.

Base de dados em ordem crescente

Salários

R\$ 12.000

R\$ 15.000

R\$ 18.000

R\$ 23.000

R\$ 25.000



**R\$ 18.000,00** é o salário **mediano** pois 50 % dos salários são maiores do que ele e 50 % dos salários são menores do que ele.

# Mediana

Cálculo da mediana para a base de dados de salários.

Como o número de observações é par a mediana é a média aritmética dos salários centrais.

## Salários

R\$ 12.000

R\$ 15.000


R\$ 18.000

R\$ 23.000

R\$ 25.000

R\$ 27.000

**R\$ 20.500,00** é o salário **mediano** pois 50 % dos salários são maiores do que ele e 50 % dos salários são menores do que ele.


$$\frac{18.000 + 23.000}{2} = 20.500$$



**economia**

A- A+ TAMANHO DA LETRA ENVIAR IMPRIMIR CORRIGIR

(0) Comentários

Votação: ★ ★ ★ ★ ★

g+ 0

Compartilhe: f t + 0

Banco Central »

# Pesquisa Focus: mercado projeta IPCA de 6,54% em 2015; previsão de PIB para 2014 diminui

Agência Estado

Publicação: 22/12/2014 09:13 Atualização:

Na penúltima divulgação do Relatório de Mercado Focus de 2014, a mediana das projeções para o IPCA de 2014 ficou estacionada em 6,38%, segundo divulgação feita nesta segunda-feira (22) pelo Banco Central. Há um mês, a taxa mediana para esse indicador estava em 6,43%.

# Percentil

## Exemplo

Suponha que um executivo de vendas deseja obter uma análise das vendas de um determinado mês realizadas por sua rede de lojas.

O analista de dados calculou o percentil 10 % e obteve o valor : 8 milhões de reais;

Dessa forma tem-se que como o percentil 10 % é 8 milhões de reais, 10 % das lojas da rede venderam menos do que 8 milhões de reais.

# Como obter o Percentil 10 % ?

# Percentil

O percentil 10 % é obtido de tal forma que 10 % das observações são menores do que ele;

O percentil  $p$  % é obtido de tal forma que  $p$  % das observações são menores do que ele;

Para se obter o percentil deve-se inicialmente ordenar as observações em ordem crescente e obter a posição do percentil.

O percentil  $p$  % é o valor que ocupa a posição :  $(n-1)\frac{p}{100} + 1$

# Percentil

## Exemplo:

Obter o percentil 10 % das  $n=9$  observações apresentadas a seguir.

Dados ordenados 30,3 31,0 31,1 31,2 31,3 31,3 31,3 31,5 31,8

$$\text{posição do percentil} = (n-1) \frac{p}{100} + 1 \quad \rightarrow \quad \text{posição do percentil 10\%} = (9-1) \frac{10}{100} + 1 = 1,8$$

O percentil 10 % ocupa a **posição 1,8**. Como não existe observação 1,8 e sim a observação 1 e a observação 2, o percentil 10 % será obtido como **a média ponderada entre as observações 1 e 2**.

# Percentil

## Exemplo:

Como o percentil 10 % será obtido como a média ponderada entre as observações 1 e 2. Deve-se definir os pesos para a observação 1 e para a observação 2.

Quando a posição de um percentil não for um número inteiro, **a parte decimal será considerada como peso.**

Como a casa decimal é 0,8 um peso será **0,8** e o outro peso **0,2**.

Como a posição é 1,8. A número 1,8 está mais próximo do número 2 do que do número 1. Desta forma, deve-se multiplicar pelo peso maior o número que está mais próximo.

No exemplo, a observação 1 será multiplicada pelo peso **0,2** e a observação 2 será multiplicada pelo peso **0,8**.



# Percentil

## Exemplo:

Dados ordenados 30,3 31,0 31,1 31,2 31,3 31,3 31,3 31,5 31,8

$$\text{Percentil 10 \%} = (0,8) * (31,0) + (0,2) * (30,3) = 30,8$$



observação 2



observação 1

O **percentil 10 %** , nesse exemplo, foi **30,8**. Desta forma tem-se que 10 % das observações são menores do que 30,8.

# Percentil

## Exemplo:

Obter o percentil 30 % das  $n=9$  observações apresentadas a seguir.

Dados ordenados 30,3 31,0 31,1 31,2 31,3 31,3 31,3 31,5 31,8

$$\text{posição do percentil} = (n-1) \frac{p}{100} + 1 \quad \Rightarrow \quad \text{posição do percentil } 30\% = (9-1) \frac{30}{100} + 1 = 3,4$$

O percentil 30 % ocupa a **posição 3,4**. Como não existe observação 3,4 e sim a observação 3 e a observação 4, o percentil 30 % será obtido como **a média ponderada entre as observações 3 e 4**.

# Percentil

## Exemplo:

Dados ordenados 30,3 31,0 31,1 31,2 31,3 31,3 31,3 31,5 31,8

Como o percentil 30 % será obtido como a média ponderada entre as observações 3 e 4. Deve-se definir os pesos para a observação 3 e para a observação 4.

Quando a posição de um percentil não for um número inteiro, **a parte decimal será considerada como peso.**

No exemplo, a observação 3 terá peso **0,6** e a observação 4 terá peso **0,4**.

Como a posição é 3,4. O número 3,4 está mais próximo do número 3 do que do número 4. Desta forma, deve-se multiplicar pelo peso maior o número que está mais próximo.

No exemplo, a observação 3 será multiplicada pelo peso **0,6** e a observação 4 será multiplicada pelo peso **0,4**.

# Percentil

## Exemplo:

Dados ordenados 30,3 31,0 31,1 31,2 31,3 31,3 31,3 31,5 31,8

$$\text{Percentil 30 \%} = (0,6) * (31,1) + (0,4) * (31,2) = 31,14$$



observação 3



observação 4

O percentil 30 % , nesse exemplo, foi 31,14. Desta forma tem-se que 30 % das observações são menores do que 31,14.

# Quartil

## Exemplo

Suponha que um executivo de vendas deseja obter uma análise das vendas de um determinado mês realizadas por sua rede de lojas.

O analista de dados calculou o primeiro quartil das vendas e obteve o valor : 15 milhões de reais;

Dessa forma tem-se que como o primeiro quartil é 15 milhões de reais, **25 % das lojas da rede venderam menos do que 15 milhões de reais.**

## Exemplo

Suponha que um executivo de vendas deseja obter uma análise das vendas de um determinado mês realizadas por sua rede de lojas.

O analista de dados calculou o terceiro quartil das vendas e obteve o valor : 45 milhões de reais;

Dessa forma tem-se que como o terceiro quartil é 45 milhões de reais, **75 % das lojas da rede venderam menos do que 45 milhões de reais.**



## **Primeiro quartil ( Q1 )**

Percentil 25 % - valor da amostra tal que 25 % das observações são menores do que ele;

## **Segundo quartil ( Q2 )**

Percentil 50 % - valor da amostra tal que 50 % das observações são menores do que ele. Este quartil também é denominado mediana;

## **Terceiro quartil ( Q3 )**

Percentil 75 % - valor da amostra tal que 75 % das observações são menores do que ele;

# Valor Discrepante

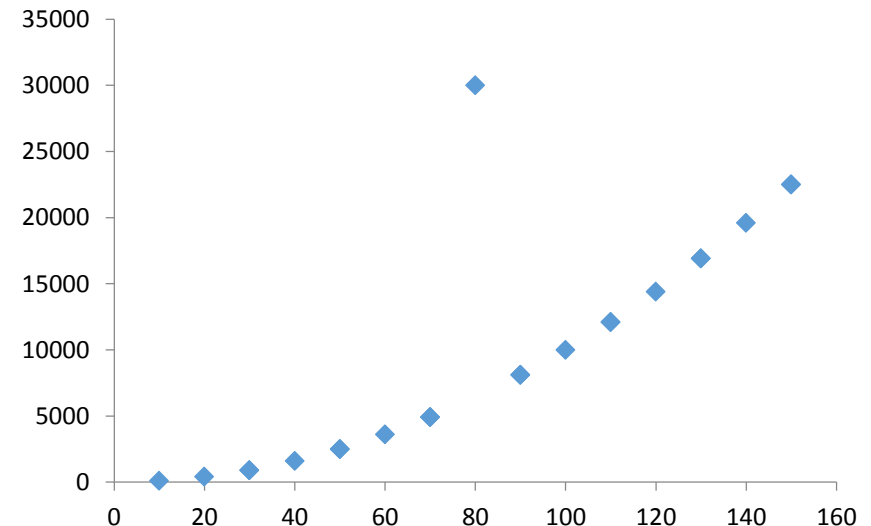
Como saber se em sua base de dados existe alguma observação muito diferente das demais ?



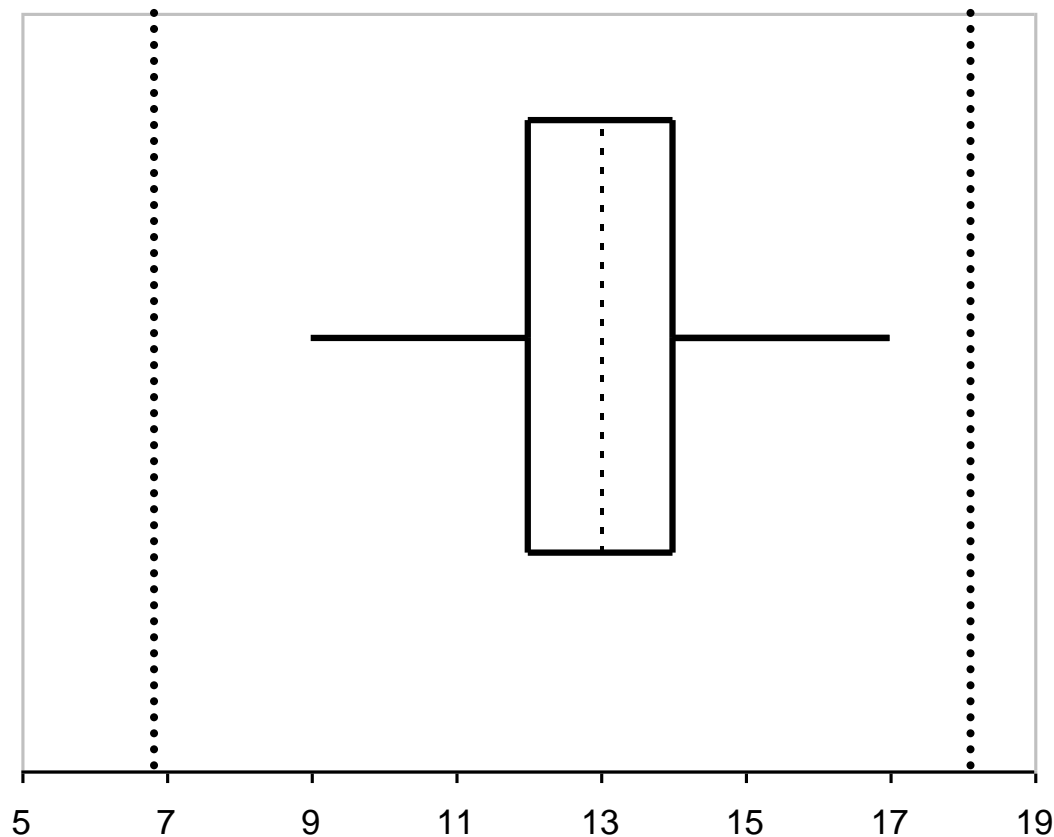
As observações muito diferente das demais são denominadas ponto fora da curva ou OUTLIER.



www.shutterstock.com · 102422164

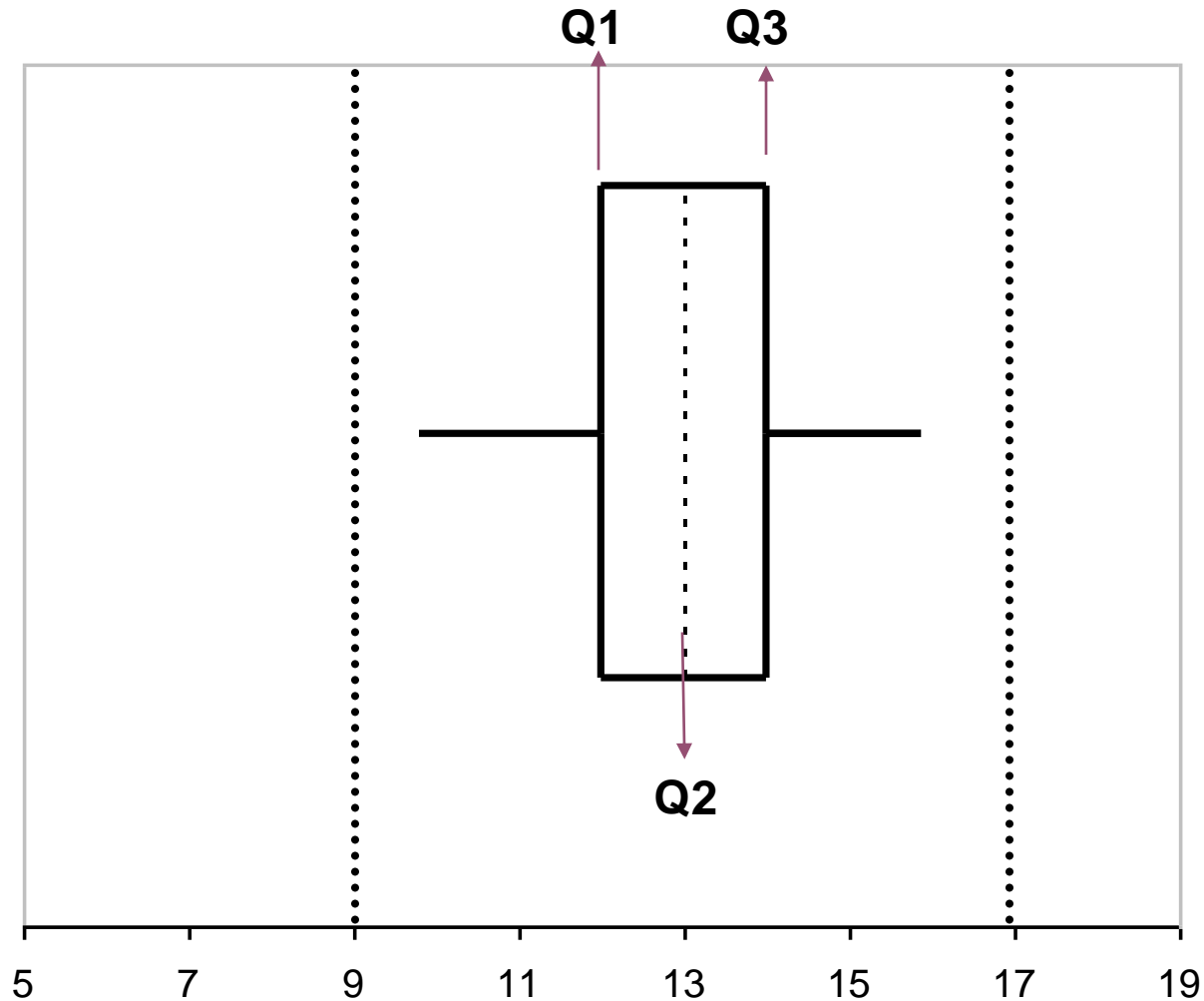


O Gráfico apresentado é denominado Box-plot. O objetivo do Box-plot é determinar se existe na base de dados alguma observação muito diferentes das demais (OUTLIER).

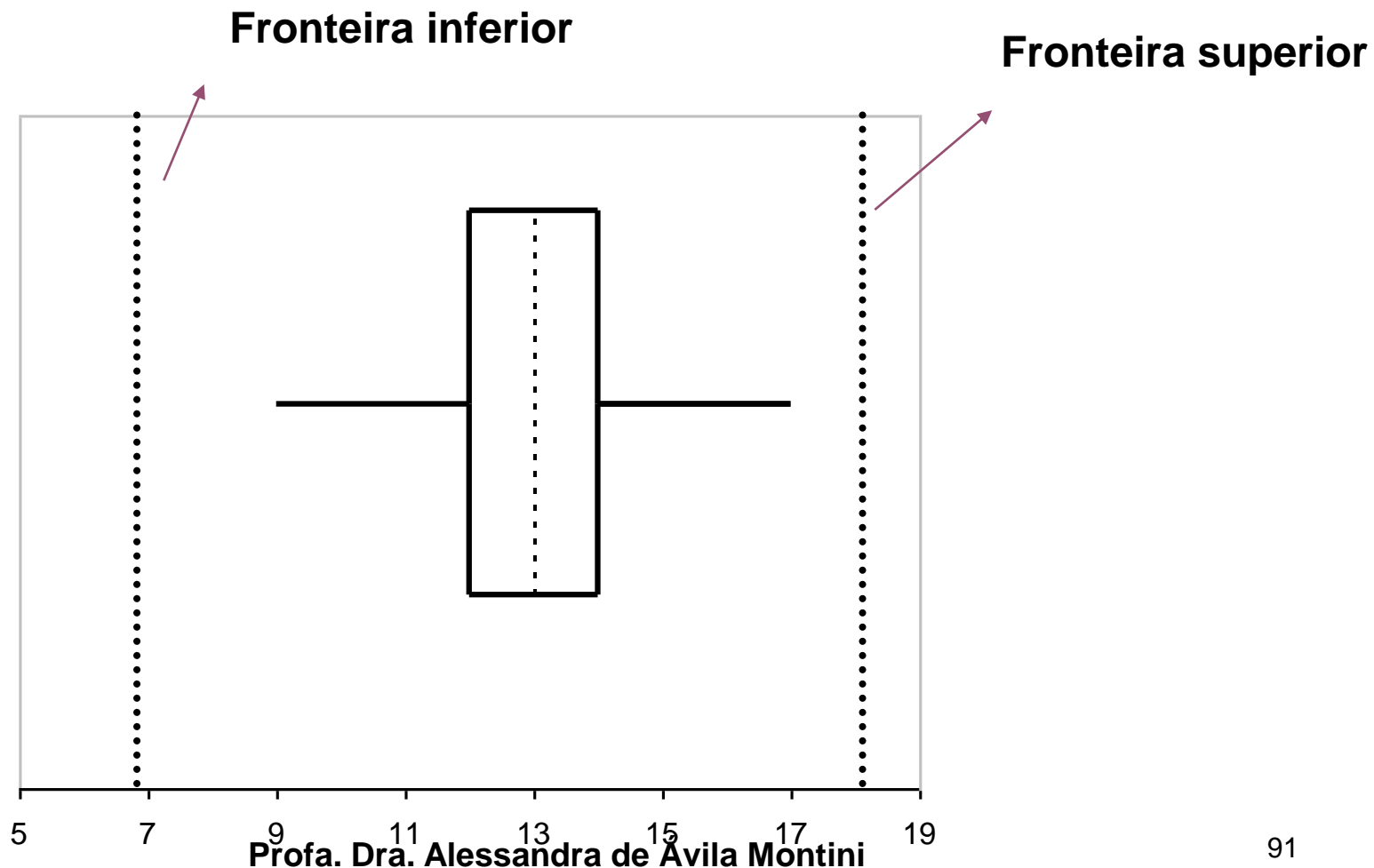


# Box-Plot

O primeiro quartil (Q1), segundo quartil (Q2) e terceiro quartil (Q3) são apresentados no Box-plot .



A fronteira inferior e a fronteira superior do Box-plot aparecem pontilhadas.

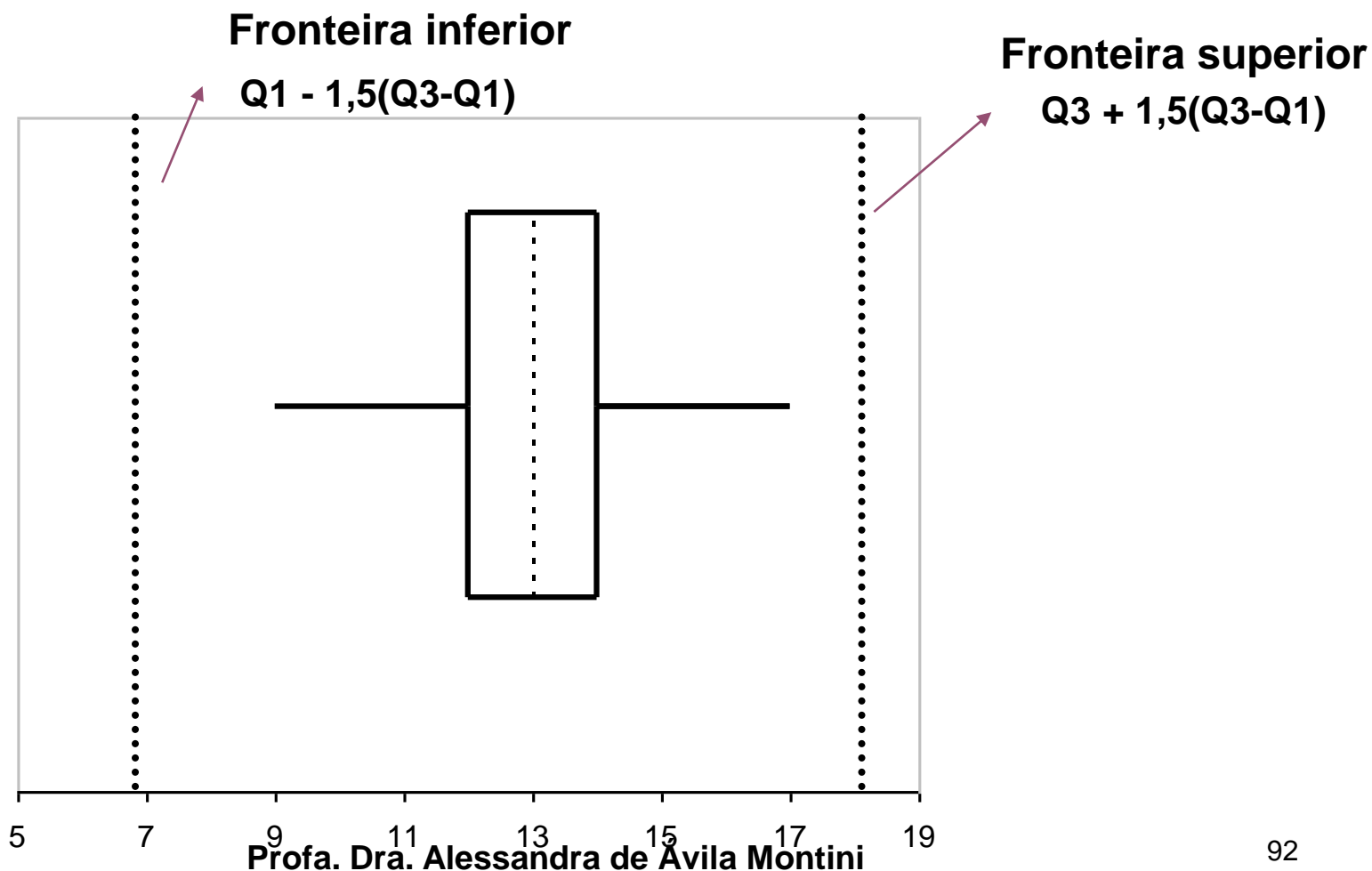




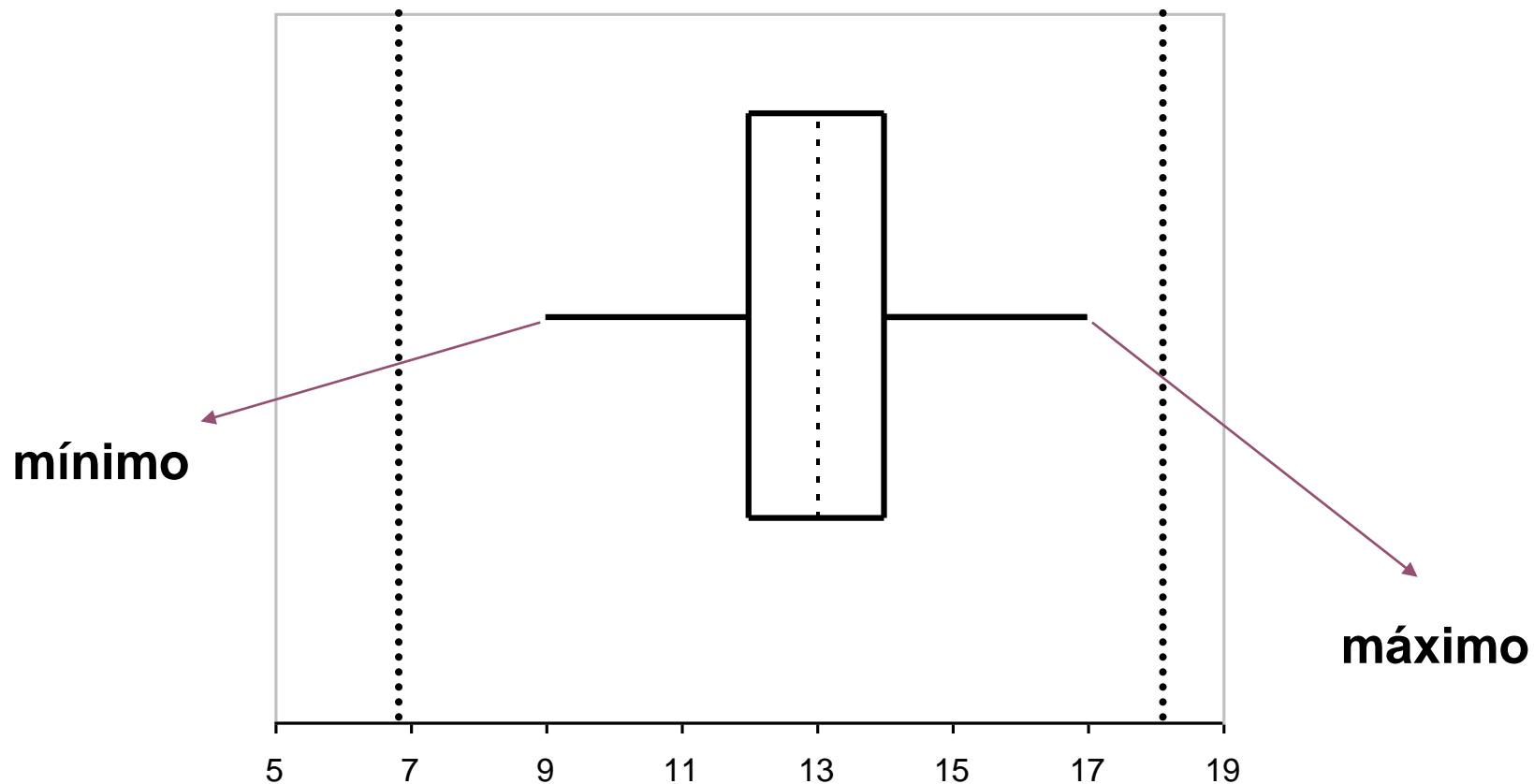
# Valor Discrepante

A fronteira inferior é dada por:  $Q1 - 1,5(Q3-Q1)$

A fronteira superior é dada por:  $Q3 + 1,5(Q3-Q1)$

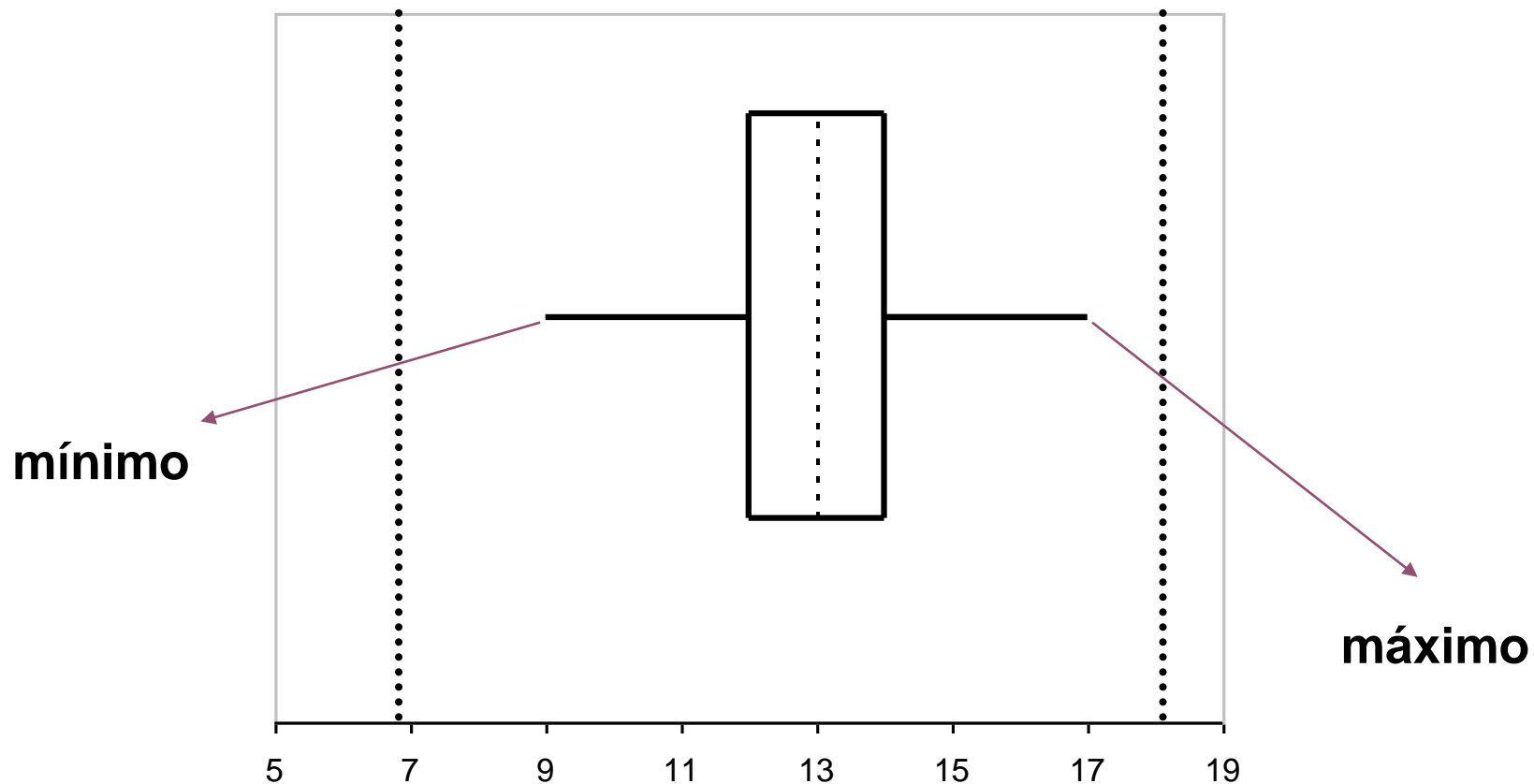


O valor **mínimo** do conjunto de observações e valor **máximo** do conjunto de observações são destacados no Box-plot.



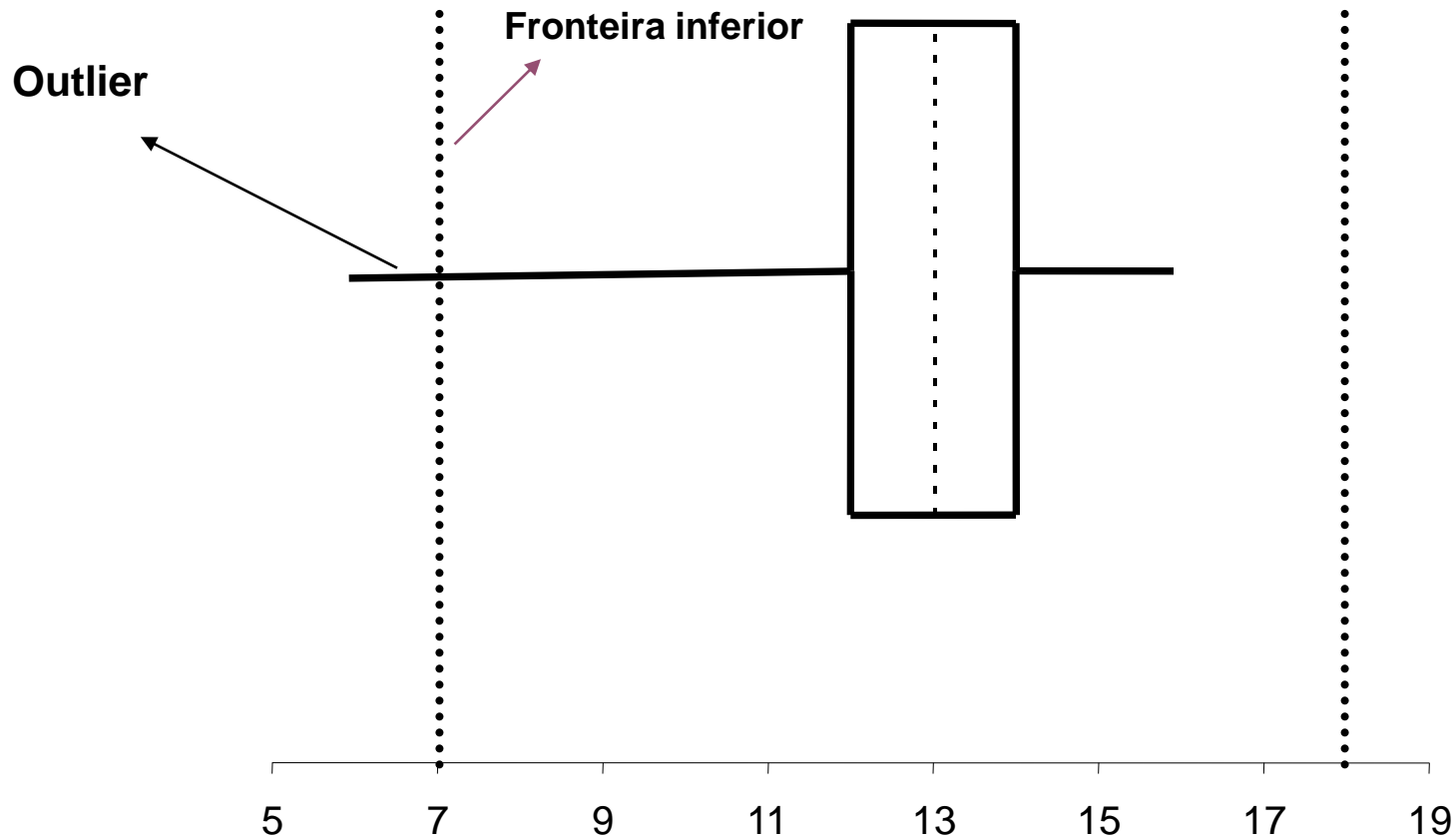
# Valor Discrepante

Quando o valor **mínimo** for superior a fronteira inferior e o valor **máximo** for inferior a fronteira superior **não existe OUTLIER**, ou seja, não existe nenhuma observação fora do padrão.



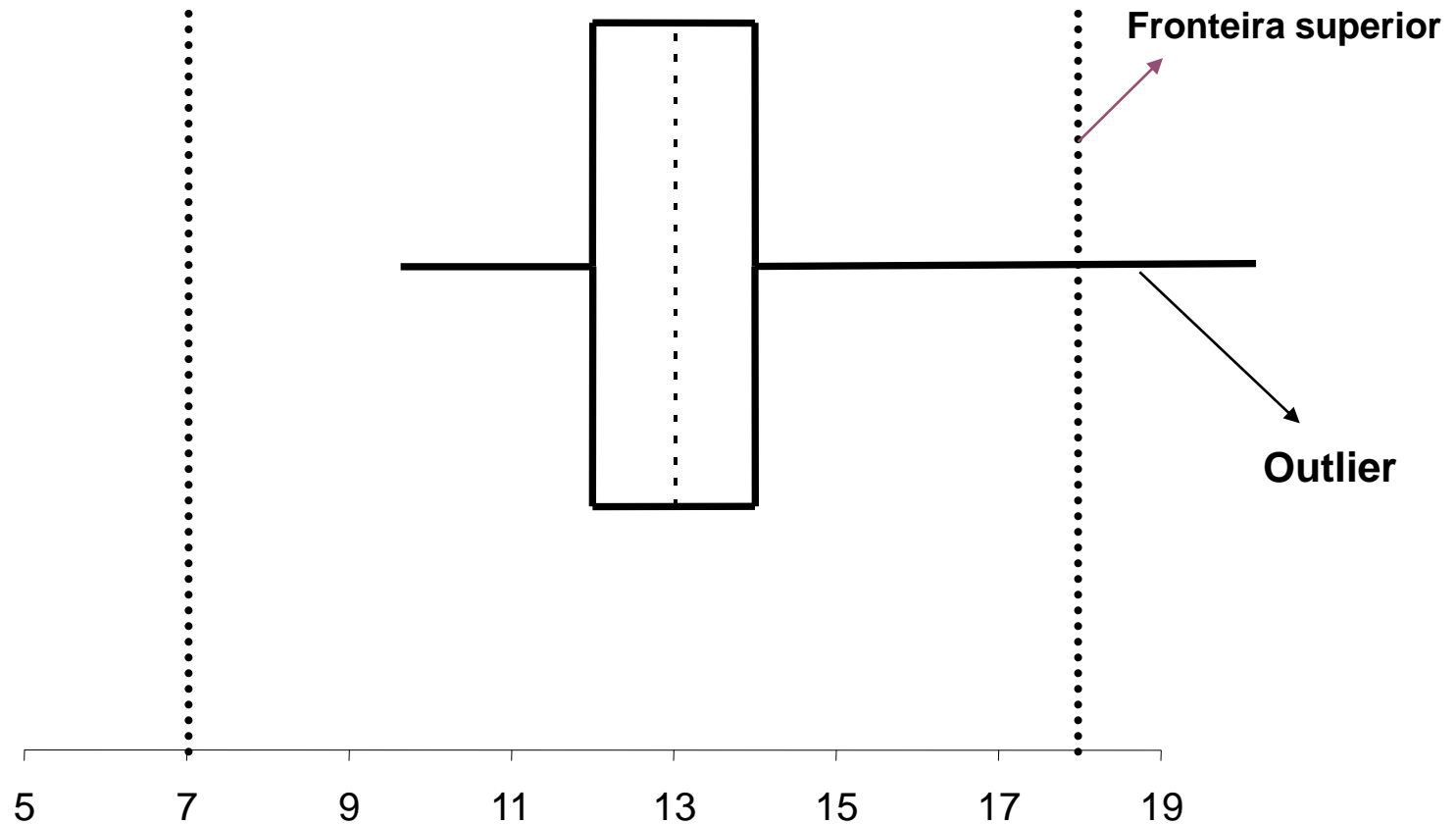
# Box-Plot

Quando o valor **mínimo** for inferior a fronteira inferior existe OUTLIER, ou seja, **existe uma ou mais observações fora do padrão**. Todas as observações inferiores a fronteira inferior são denominadas outlier.

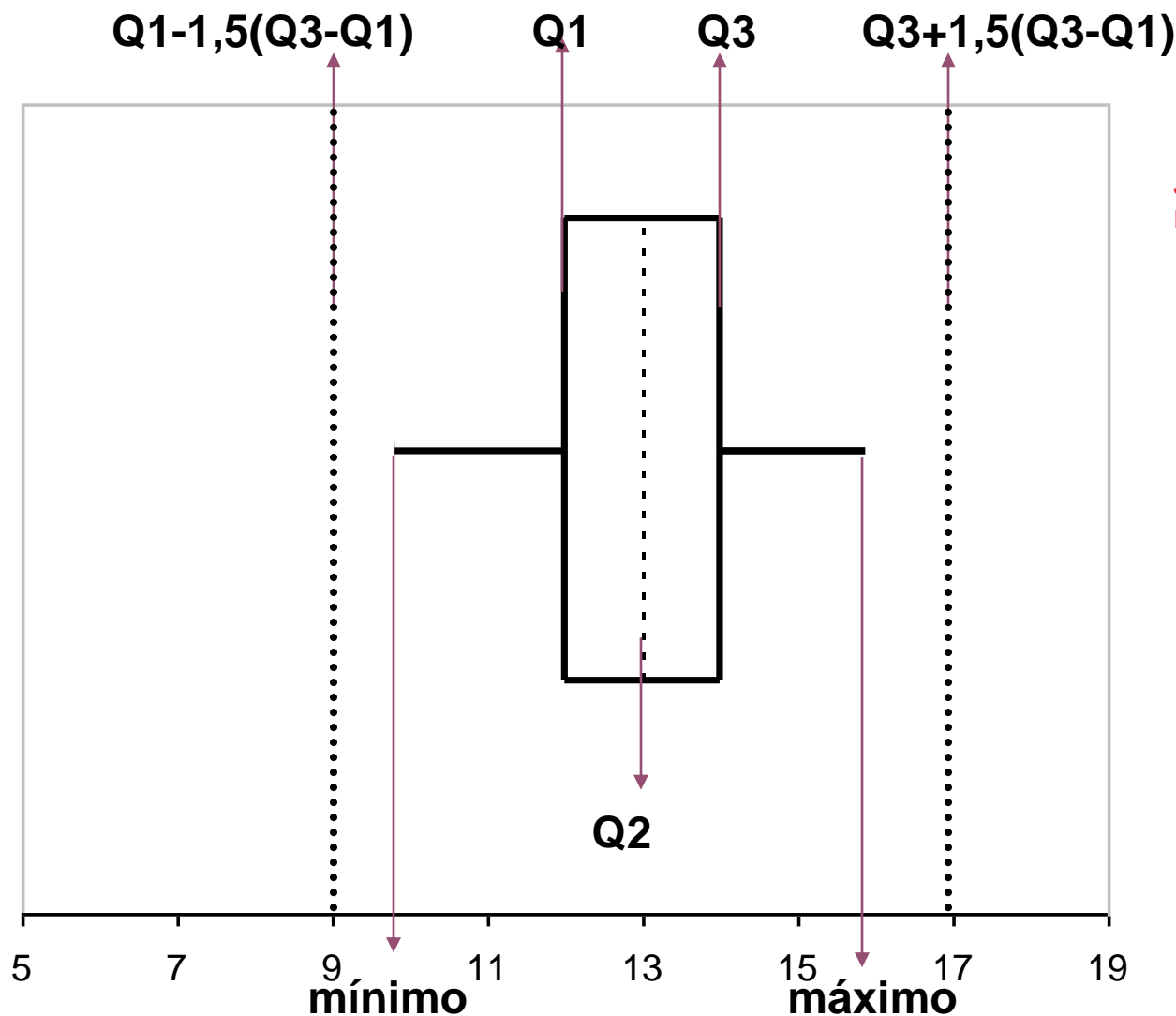


# Box-Plot

Quando o valor **máximo** for superior a fronteira superior existe OUTLIER, ou seja, existe uma ou mais observações fora do padrão. Todas as observações superiores a fronteira superior são denominadas outlier.



O Box-plot contém as fronteiras que aparecem pontilhadas e deve-se ter atenção ao mínimo, máximo, primeiro quartil (Q1), segundo quartil (Q2) e terceiro quartil (Q3).



John Tukey - Outliers  
Exploratory Data Analysis

# Mínimo e Máximo



# Mínimo e Máximo

O **mínimo** é o menor valor do conjunto de observações.

O **máximo** é o maior valor do conjunto de observações.

## Salários

R\$ 12.000



Salário **mínimo**

R\$ 15.000

R\$ 18.000

R\$ 23.000

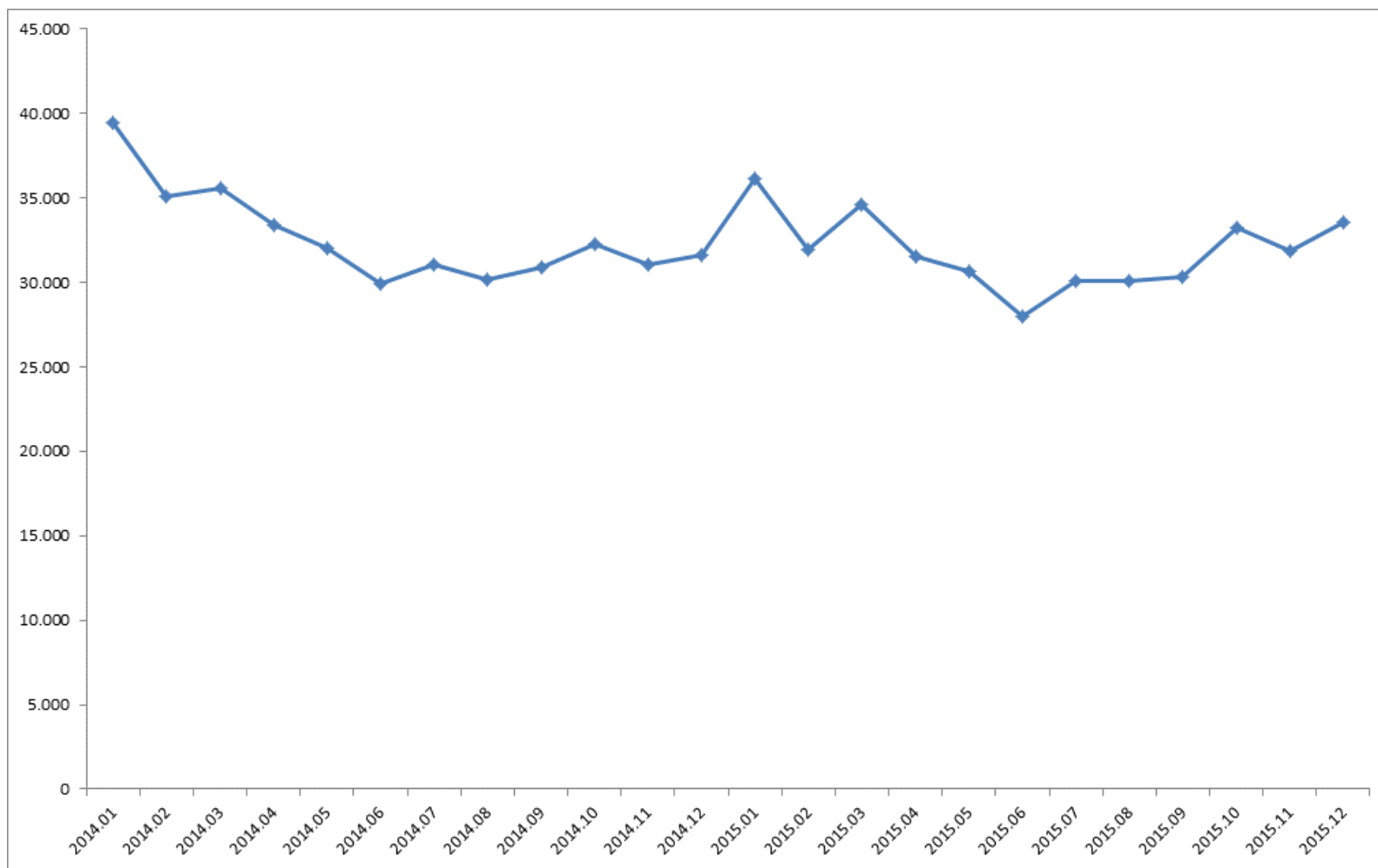
R\$ 25.000



Salário **máximo**

# Exercício

## Geração de energia elétrica - hidráulica - qde. - GWh - 2014 e 2015



**Fonte: Operador Nacional do Sistema Elétrico (ONS) - ONS12\_HIDR12**

## Geração de energia elétrica - hidráulica - qde. – GWh – 2014 e 2015

Obtenha para cada ano: Média, Mediana, Q1, Q3 e Box-plot

2014.01	39.433	2015.01	36.175
2014.02	35.095	2015.02	31.901
2014.03	35.609	2015.03	34.626
2014.04	33.411	2015.04	31.506
2014.05	31.988	2015.05	30.652
2014.06	29.935	2015.06	28.018
2014.07	31.039	2015.07	30.047
2014.08	30.179	2015.08	30.092
2014.09	30.916	2015.09	30.334
2014.10	32.281	2015.10	33.257
2014.11	31.016	2015.11	31.888
2014.12	31.597	2015.12	33.561

# Medidas de Dispersão

**Os dois grupos apresentam a mesma variabilidade ?**

<b>Homens</b>	<b>Mulheres</b>
<b>52000</b>	<b>40000</b>
<b>50000</b>	<b>50000</b>
<b>48000</b>	<b>60000</b>
<b>50000</b>	<b>50000</b>

A variabilidade no grupo das mulheres é maior do que a variabilidade no grupo dos homens.

# Desvio



# Medidas de Dispersão

O desvio é a distância de cada observação à média.

Qualquer conjunto de dados, a soma dos desvios sempre será igual à zero.

	Homens	Desvio	Mulheres	Desvio
	52000	2000	40000	-10000
	50000	0	50000	0
	48000	-2000	60000	10000
	50000	0	50000	0
Média	50000		50000	

# Desvio Médio Absoluto

O desvio médio absoluto é a média aritmética dos desvios.

	Homens	Mulheres
	52000	40000
	50000	50000
	48000	60000
	50000	50000
Média	50.000	50.000
Desvio Médio	1000	5000

O desvio médio absoluto para os homens é dado por:

$$\text{Desvio Medio Absoluto} = \frac{|2000| + |0| + |-2000| + |0|}{4} = \frac{4.000}{4} = 1.000$$

# **Variância e Desvio Padrão Populacional**

# Medidas de Dispersão

A variância populacional é a média aritmética dos desvios elevados ao quadrado.

	Homens	Desvio	Mulheres	Desvio
	52000	2000	40000	-10000
	50000	0	50000	0
	48000	-2000	60000	10000
	50000	0	50000	0
Média	50000		50000	

A variância populacional para os homens é dada por:

$$\text{Variância} = \frac{(2000)^2 + (0)^2 + (-2000)^2 + (0)^2}{4} = \frac{8.000.000}{4} = 2.000.000$$

# Medidas de Dispersão

Como a variância populacional está na unidade ao quadrado, para retornar a unidade original deve-se obter a raiz quadrada da variância populacional.

O desvio padrão populacional é a raiz quadrada da variância populacional.

$$\text{Desvio Padrão} = \sqrt{\text{variância}} = \sqrt{2.000.000} = 1.414$$

# **Variância e Desvio Padrão**

## **Amostrai**



# Medidas de Dispersão

A variância amostral é obtida por meio da soma dos desvios elevados ao quadrado dividindo-se pelo total de observações menos um.

	Homens	Desvio	Mulheres	Desvio
	52000	2000	40000	-10000
	50000	0	50000	0
	48000	-2000	60000	10000
	50000	0	50000	0
Média	50000		50000	

A variância amostral para os homens é dada por:

$$\text{Variância} = \frac{(2000)^2 + (0)^2 + (-2000)^2 + (0)^2}{3} = \frac{8.000.000}{3} = 2.666.666,66$$

graus de liberdade ou esperança matemática

Se for população é n e se for amostra n-1

# Medidas de Dispersão

Como a variância amostral está na unidade ao quadrado, para retornar a unidade original deve-se obter a raiz quadrada da variância amostral.

O desvio padrão amostral é a raiz quadrada da variância amostral.

$$\text{Desvio Padrão} = \sqrt{\text{variância}} = \sqrt{2.666.666,66} = 1.633$$

O variância amostral e desvio padrão amostral.

	Homens	Mulheres
	52000	40000
	50000	50000
	48000	60000
	50000	50000
Média	50.000	50.000
Variância	2.666.667	66.666.667
Desvio Padrão	1.633	8.165

# Coeficiente de Variação

O Coeficiente de Variação é uma medida de dispersão relativa usada na comparação da variabilidade para dados que diferem no valor médio.

O Coeficiente de Variação é obtido por meio da divisão do desvio padrão pela média multiplicando-se por 100.

$$CV = \frac{\text{Desvio Padrão}}{\text{Média}} \times 100$$

## Exemplo

Cálculo do coeficiente de variação para a base de dados de salários.

$$CV = \frac{\text{Desvio Padrão}}{\text{Média}} \times 100$$

	Homens	Mulheres
	52000	40000
	50000	50000
	48000	60000
	50000	50000
Média	50000	50000
Desvio Padrão	1633	8165
Coef. de Variação	3.3	16.3

# Medidas de Dispersão

Quando as médias dos grupos a serem comparados diferem a utilização do coeficiente de variação é muito importante para a comparação entre as variabilidades dos grupos.

O Coeficiente de Variação é obtido por meio da divisão do desvio padrão pela média multiplicando-se por 100.

O grupo com o maior Coeficiente de Variação pode ser considerado o grupo com maior variabilidade.

Média	2000	4000	10000	20000	50000
Desvio Padrão	200	200	200	200	200
Coeficiente de Variação	10.00	5.00	2.00	1.00	0.40



**Maior Variabilidade**



**Menor Variabilidade**



# Exercício

Rendimento em R\$ (mil)

Mês	Fundo 1	Fundo 2	Fundo 3
Janeiro	14	30	5
Fevereiro	13	26	20
Março	16	14	8

Em qual dos 3 fundos de investimentos apresenta a maior variabilidade ?



Mês	Fundo 1	Fundo 2	Fundo 3
Janeiro	14	30	5
Fevereiro	13	26	20
Março	16	14	8
Média	14,3	23,3	11,0
Desvio Padrão	1,5	8,3	7,9
Coeficiente de Variação	10,7	35,7	72,2

# Exercício

## Geração de energia elétrica - hidráulica - qde. – GWh – 2014 e 2015

Calcule para cada ano: Desvio Médio, Desvio Padrão e Coeficiente de Variação.

2014.01	39.433	2015.01	36.175
2014.02	35.095	2015.02	31.901
2014.03	35.609	2015.03	34.626
2014.04	33.411	2015.04	31.506
2014.05	31.988	2015.05	30.652
2014.06	29.935	2015.06	28.018
2014.07	31.039	2015.07	30.047
2014.08	30.179	2015.08	30.092
2014.09	30.916	2015.09	30.334
2014.10	32.281	2015.10	33.257
2014.11	31.016	2015.11	31.888
2014.12	31.597	2015.12	33.561

# Coeficiente de Assimetria

# Coefficiente de Assimetria

## Coefficiente de Assimetria

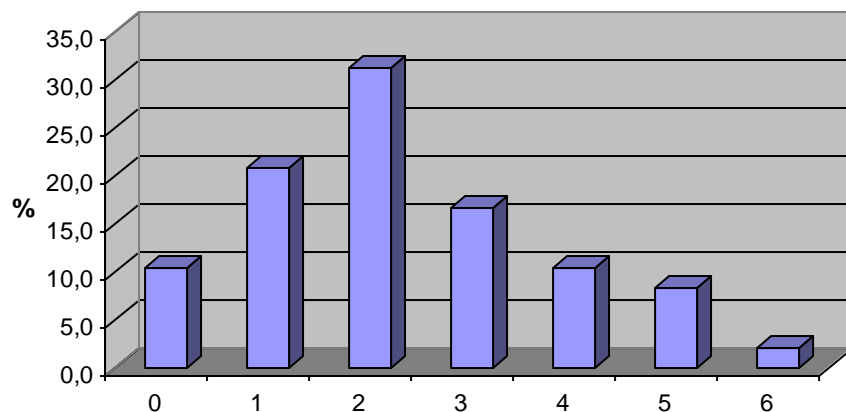
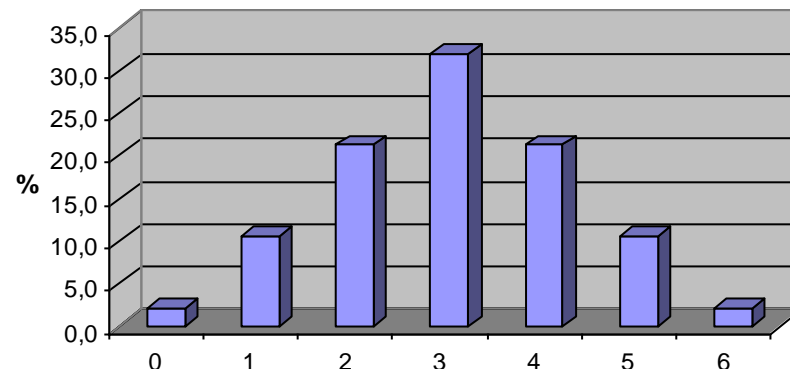
$$A = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{S} \right)^3$$

$n$  : tamanho da amostra

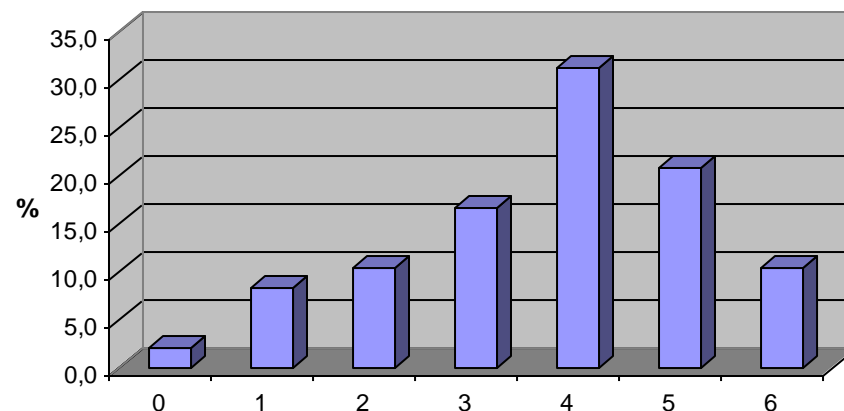
$S$  : desvio padrão amostral

$\bar{X}$  : média amostral

**Distribuição Simétrica**  
**A=0**



**A > 0 – Distribuição inclinada para a direita**  
**Distribuição Assimétrica à Direita**



**A < 0 - Distribuição inclinada para a esquerda**  
**Distribuição Assimétrica à Esquerda**

# HP 12 C

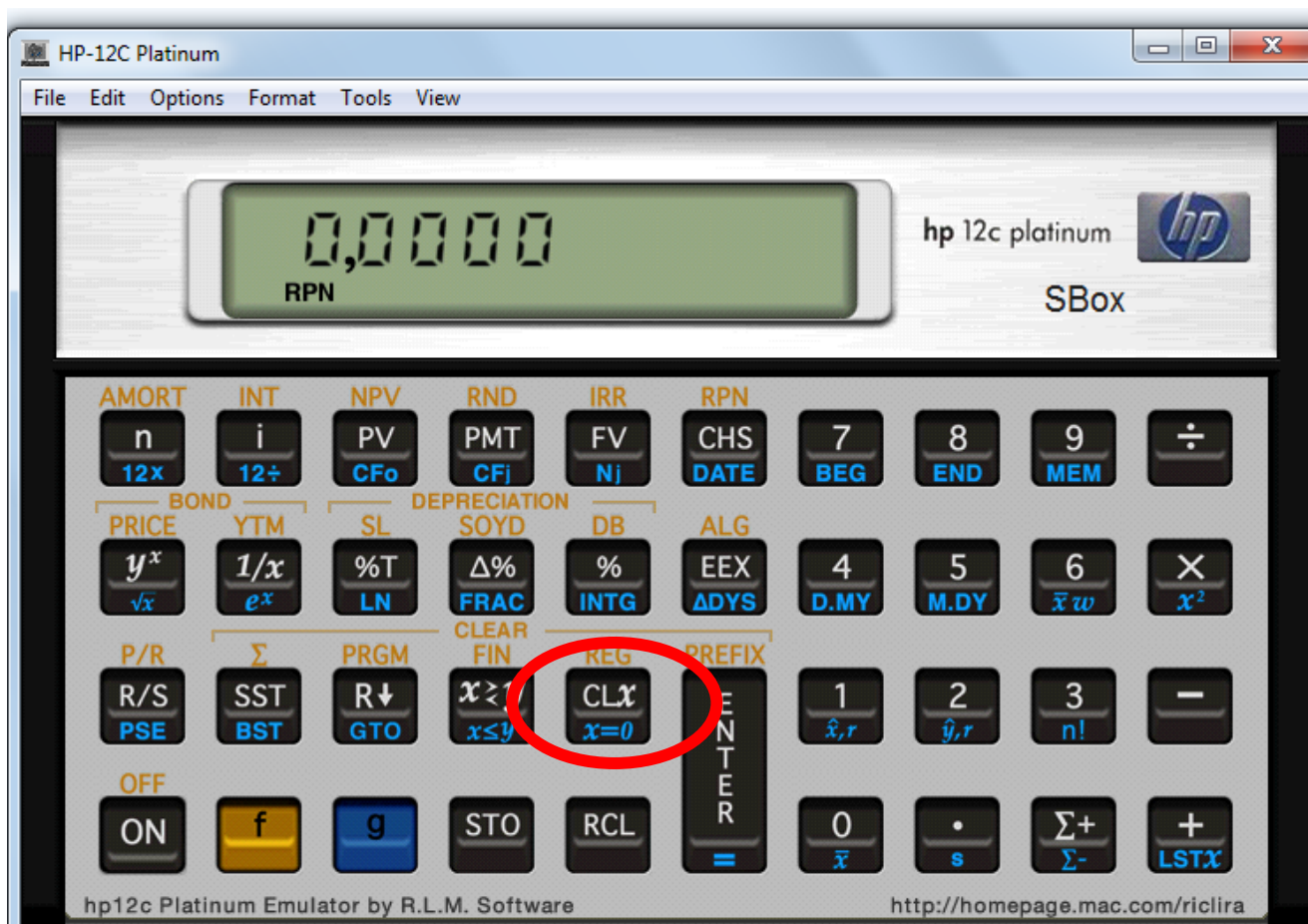


# HP 12C

Para se obter a média e o desvio padrão amostral na HP 12C deve seguir os passos:

1 – Apagar as observações armazenadas na HP por meio da tecla

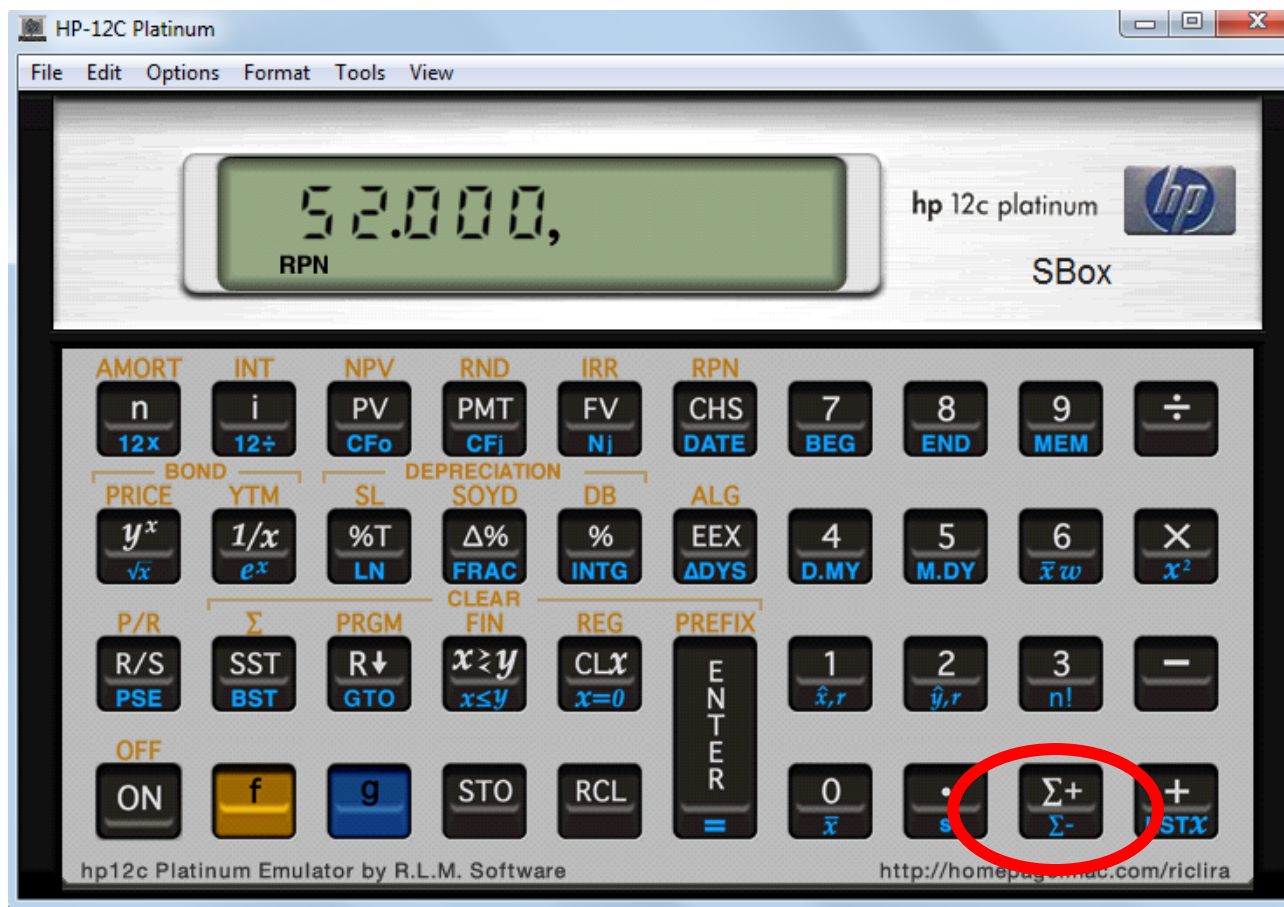
**CLX**



# HP 12C

2 – Armazenar as observações : **52.000** ; **50.000** ; **48.000** ; **50.000** por meio do procedimento:

- Digitar 52000 e clicar na tecla  $\Sigma+$



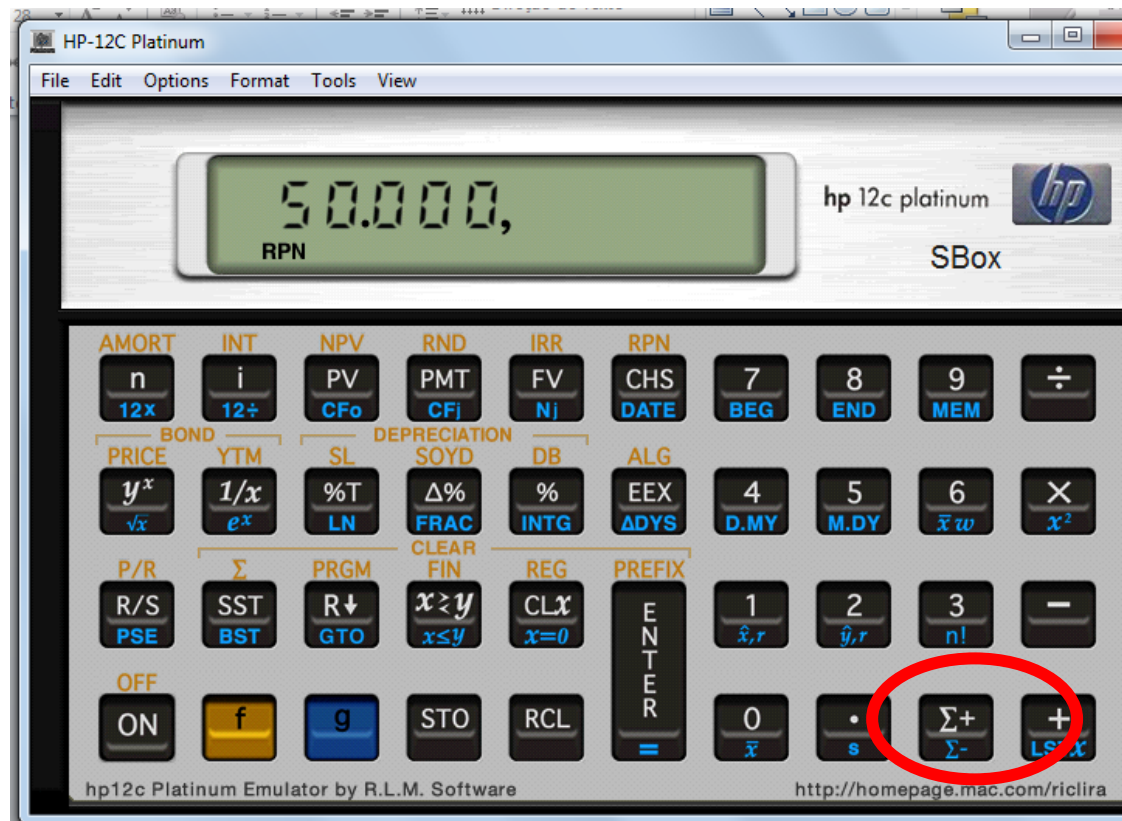
# HP 12C

Aparecerá o número 1 no visor da HP indicando que a primeira observação foi armazenada



# HP 12C

- Digitar 50.000 e clicar na tecla  $\Sigma+$

 $\Sigma+$ 

Aparecerá o número 2 no visor da HP indicando que a segunda observação foi armazenada.

# HP 12C

- Digitar 48.000 e clicar na tecla

A black rectangular button with a white border, containing the white text  $\Sigma+$ .

Aparecerá o número 3 no visor da HP indicando que a terceira observação foi armazenada.

- Digitar 50.000 e clicar na tecla

A black rectangular button with a white border, containing the white text  $\Sigma+$ .

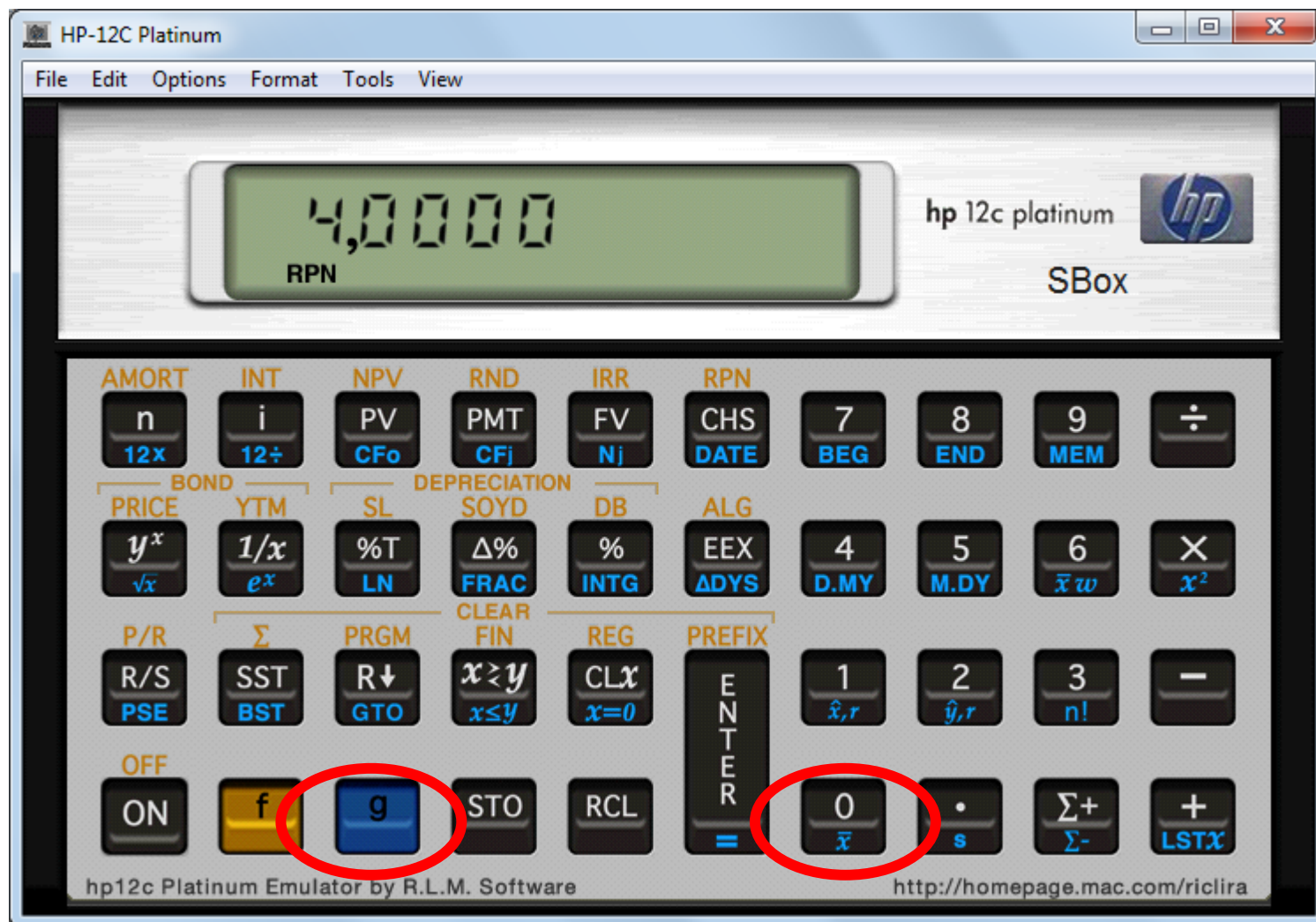
Aparecerá o número 4 no visor da HP indicando que a quarta observação foi armazenada.

# HP 12C

- Para obter a média clicar na tecla

**g**

e depois na tecla

**0**



# HP 12C

A média para esse conjunto de observações é 50.0000,00



# HP 12C

- Para obter o desvio padrão AMOSTRAL clicar na tecla **g** e na tecla **.**

