Dados

Tipos de Dados

- Os conjuntos de dados são compostos por dois tipos de dados: dados **qualitativos** e dados **quantitativos**

Dados

Qualitativos	Quantitativos
Consiste de atributos, rótulos ou entidades não numéricas	Consiste de medições numéricas ou contagens
(Categóricos)	(Numéricos)

Dados

Tipos de Dados

- Considere o seguinte exemplo
  - Notas de alunos em uma determinada disciplina

Aluno	Nota
Sally	3.22
Bob	3.98
Cindy	2.75
Mark	2.24
Kathy	3.84

Qualitativo

Quantitativo

Dados

Escalas de Medição

- Há quatro escalas de medição de um dado, quais sejam:

Escalas de Medição

Nominal

Ordinal

Intervalar

Razão

Aumento do número de operações

Dados

Escala Nominal

- Representam **categorias** que não mantém necessariamente relação entre elas
- Não é possível realizar operações aritméticas (soma, média, etc.)
- Normalmente realiza-se apenas a contagem das observações em cada categoria

Escalas de Medição

Nominal

Calculada usando nomes, rótulos ou categorias. Não há operações aritméticas feitas nesta escala.

Cores da bandeira do Brasil

Nomes dos alunos em uma classe

Livros que estamos lendo

Dados

Escala Ordinal

- **Categorias** podem ser representadas por nomes, símbolos ou números, porém há uma ordenação de uma categoria em relação à outra
- A distância entre uma categoria e a outra não pode ser medida numericamente
- Além da operação de contagem, permitem operações que envolvam ordenação (maior/menor)

Escalas de Medição → **Ordinal**

Podem ser ordenadas entre si, mas não é possível diferenciá-las numericamente.

Nível de experiência: junior, pleno e senior	Números das camisas dos jogadores da seleção	Top 10 músicas mais tocadas no momento

Dados

Escala Intervalar

- Escala **quantitativa**
- O valor nulo não corresponde à ausência da característica medida
- A escala possui um zero arbitrário
- Exemplo: 0 **°C** não significa ausência de temperatura (-273 **°C**)
- Operação de divisão é ilegítima em dados intervalares

Escalas de Medição → **Intervalar**

Ordenadas, a diferença entre duas medidas pode ser calculada

Temperatura

Linha do tempo em anos

40 ° não é o dobro de 20 °C.

Veja em Fahrenheit:

20 °C → 68°F

40 °C → 104 °F

Dados

Escala Razão

- Escala **quantitativa**
- O zero corresponde à ausência da característica medida
- É possível realizar todas as operações aritméticas em dados dessa escala

Escalas de Medição

Razão

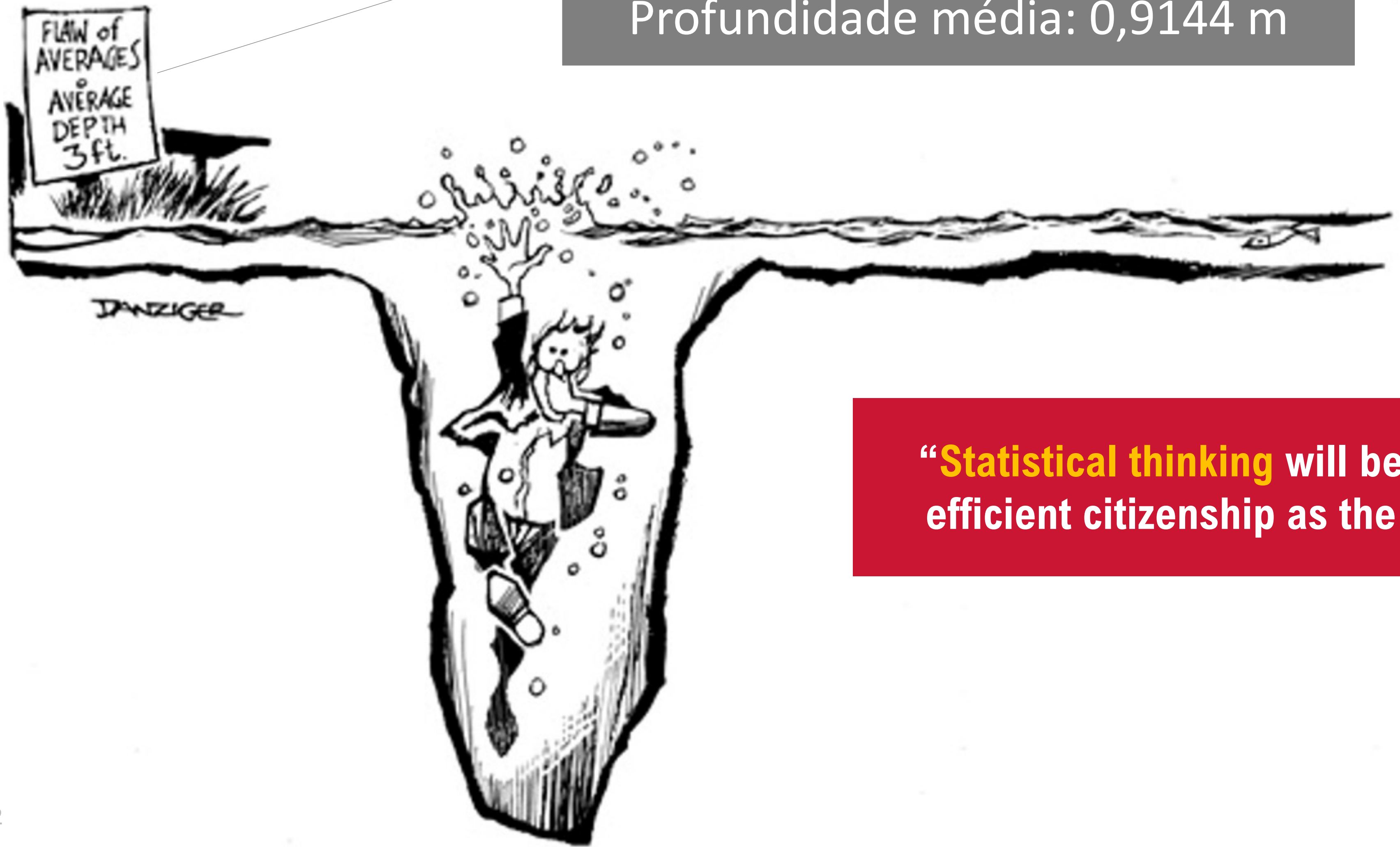
Uma divisão entre dois valores pode ser realizada

Idade | Peso | Altura | Dinheiro

Dados Quantitativos

Descrevendo os dados numericamente - motivação

Profundidade média: 0,9144 m



"**Statistical thinking** will be one day as necessary for efficient citizenship as the ability to **read** and **write**"

Hebert George Wells

Escritor Britânico (1940)

Insper

Dados Quantitativos

Descrevendo os dados numericamente

Descrevendo os dados numericamente

- **Centralidade e Posição**
  - Média
  - Mediana
  - Moda
  - Média Ponderada
  - Percentís
  - Quartis
- **Dispersão**
  - Amplitude
  - Invervalo interquartil
  - Variância
  - Desvio Padrão

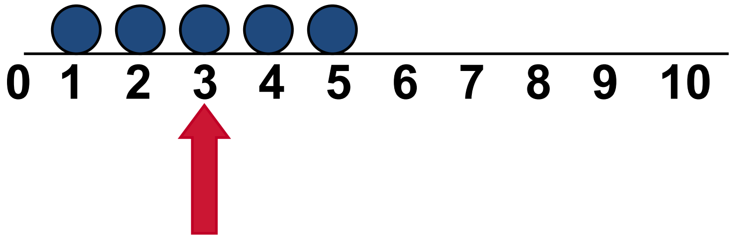
Medidas de Centralidade e Posição

Média

- A mais comum das medidas de tendência central

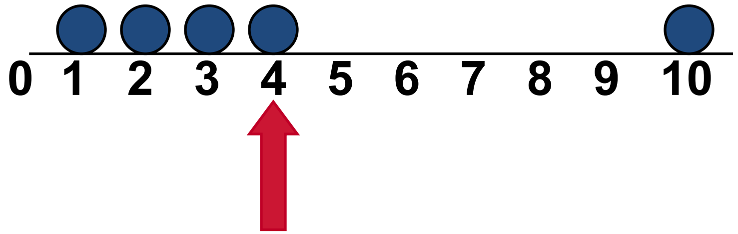
Média = soma dos valores dividida pela quantidade dos valores

- A média é afetada por valores extremos (outliers)



Média = 3

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



Média = 4

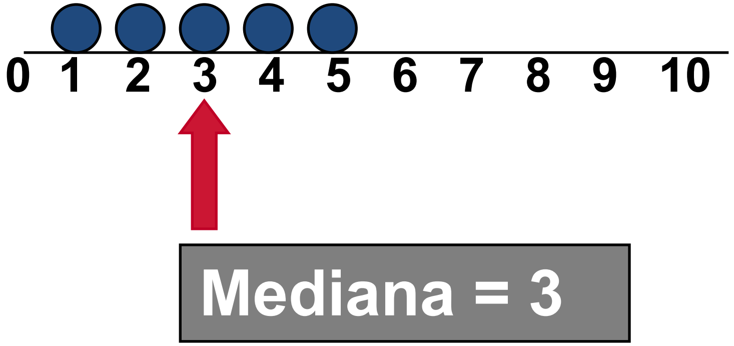
$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

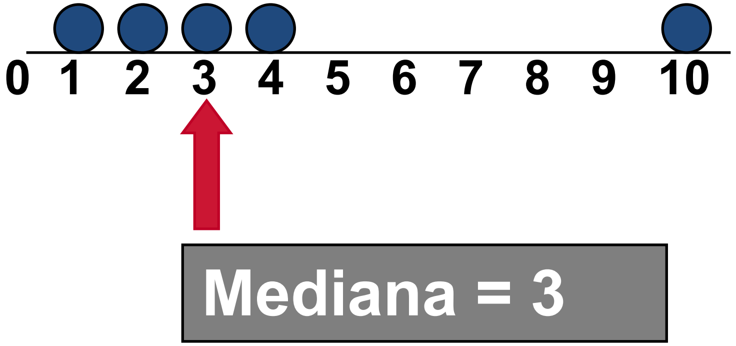
Adaptado de: Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc.

Medidas de Centralidade e Posição

Mediana

- Em um vetor ordenado a mediana é o elemento do meio
- A mediana não é afetada por valores extremos



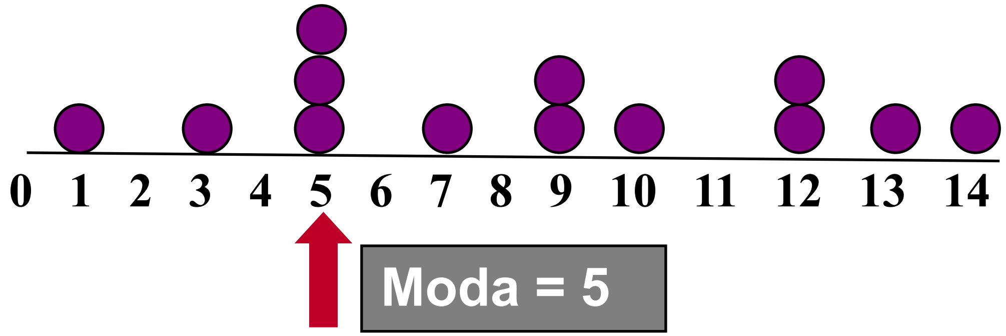


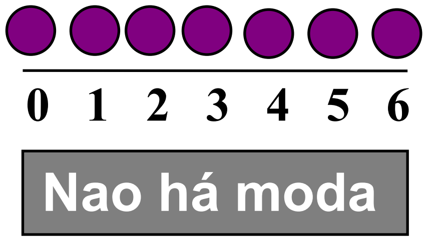
- Se o vetor é par, a mediana é a média dos elementos centrais

Medidas de Centralidade e Posição

Moda

- Representa o valor que ocorre com maior frequência no conjunto de dados
- Não é afetada por valores extremos
- Utilizada tanto por dados qualitativos quanto quantitativos
- É possível que não haja moda
- É possível também que possam existir várias modas





Adaptado de: Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc.

Medidas de Centralidade e Posição

Média ponderada

- Utilizada quando os valores estão agrupados pela frequência ou importância relativa

Exemplo: Amostra de 26 tarefas

Dias para finalização	Frequência
5	4
6	12
7	8
8	2

Média ponderada dos dias para finalização:

$$\bar{X}_W = \frac{\sum w_i x_i}{\sum w_i} = \frac{(4 \times 5) + (12 \times 6) + (8 \times 7) + (2 \times 8)}{4 + 12 + 8 + 2}$$

$$= \frac{164}{26} = 6.31$$

Adaptado de: Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc.

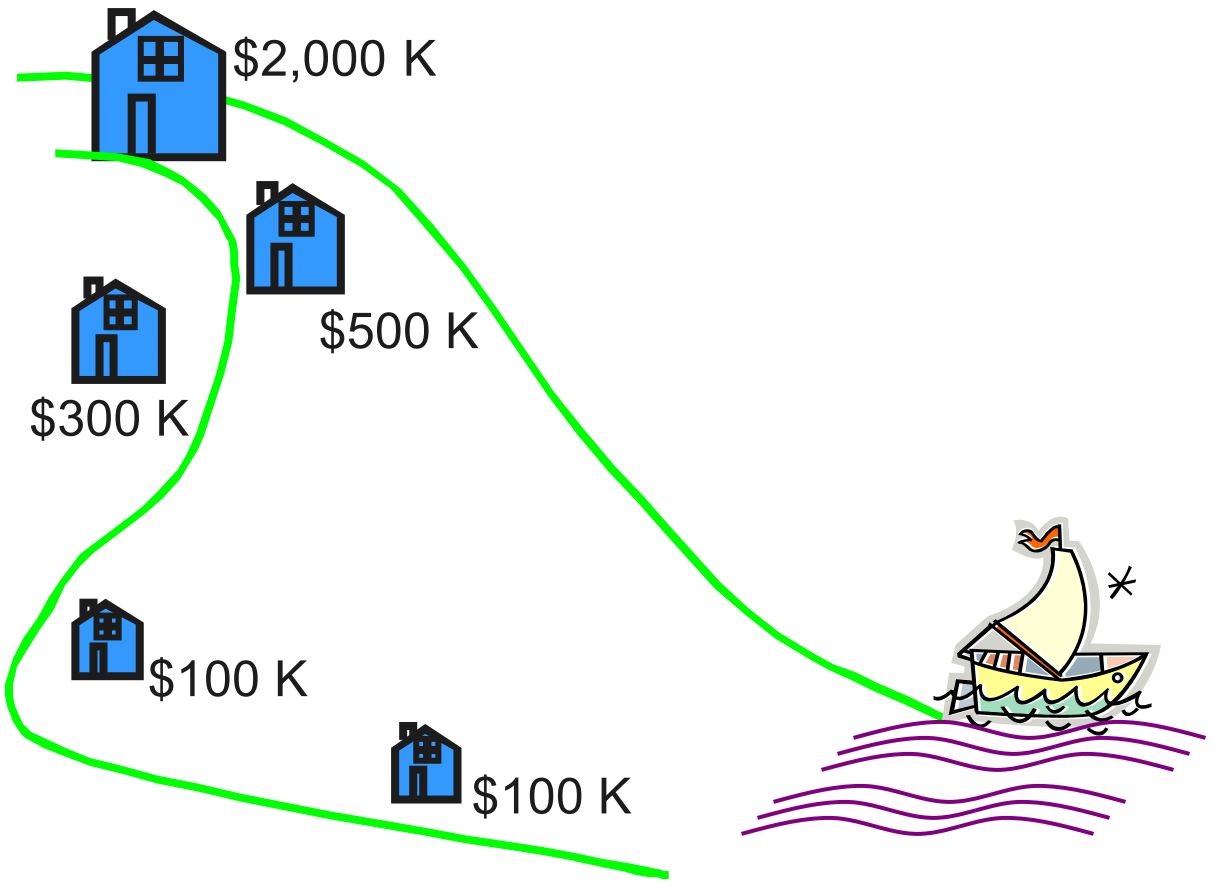
Exemplo

Casas de veraneio

- Cinco casas de veraneio

Preços:

$2,000,000
500,000
300,000
100,000
100,000



Insper

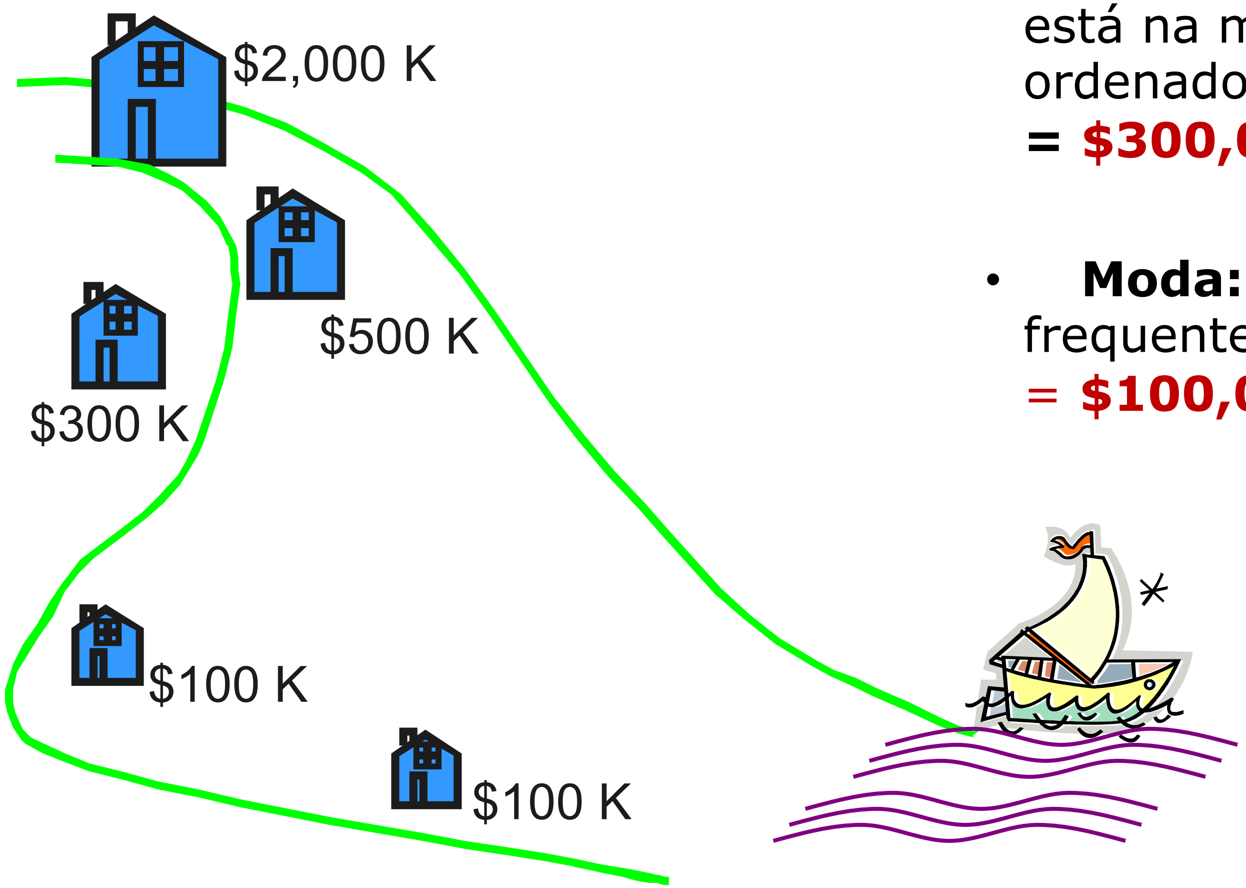
Exemplo

Casas de veraneio

- Cinco casas de veraneio

Preços:

$2,000,000
500,000
300,000
100,000
100,000



- **Média**: (3,000,000 / 5) = **$600,000**

- **Mediana:** valor que está na metade do vetor ordenado. **= $300,000**

- **Moda:** valor mais frequente = **$100,000**

Adaptado de: Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc.

Insper

Medidas de Centralidade e Posição

Outro exemplo

- Uma organização fundada há 29 anos se une a outras empresas mais antigas:
  - Tempo de existência das organizações:

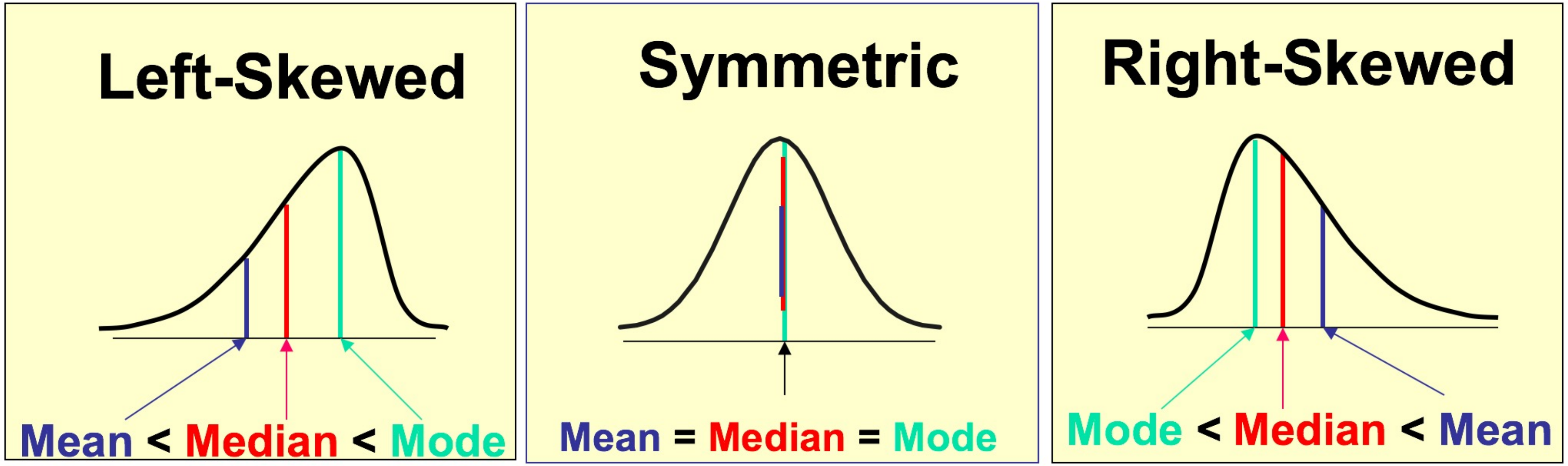
$$53\ 32\ 61\ 57\ 39\ 44\ 57\ 29$$

  - Calcule a média, a mediana e a moda. Qual medida de tendência central está sendo mais afetada pela nova empresa?

Adaptado de: Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc.

Distribuição dos Dados

- Simétrica ou Assimétrica (Skewed)



(Longer tail extends to left)

(Longer tail extends to right)

Adaptado de: Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc.

Insper

Medidas de Posição

Percentis e Quartis

Medidas de Posição

Percentis

O p percentil em um conjunto de dados significa que:

p% dados são menores ou iguais a esse valor

(100 – p)% é maior ou igual a esse valor (onde $0 \leq p \leq 100$)

Quartis

- 1° Quartil = 25 percentil
- 2° Quartil = 50 percentil = mediana
- 3° Quartil = 75 percentil

Medidas de Posição

Percentis e Quartis

- Os quartis dividem os dados em quatro subconjuntos:

25%	25%	25%	25%

Q1 Q2 Q3

- Exemplo
  - Determinar o primeiro quartil

Vetor ordenado:	**11**	**12**	**13**	**16**	**16**	**17**	**18**	**21**	**22**

(n = 9)

Q1 = 25 percentil, o qual está na posição $\frac{25}{100}(9+1) = 2.5$

$$i = \frac{p}{100}(n+1)$$

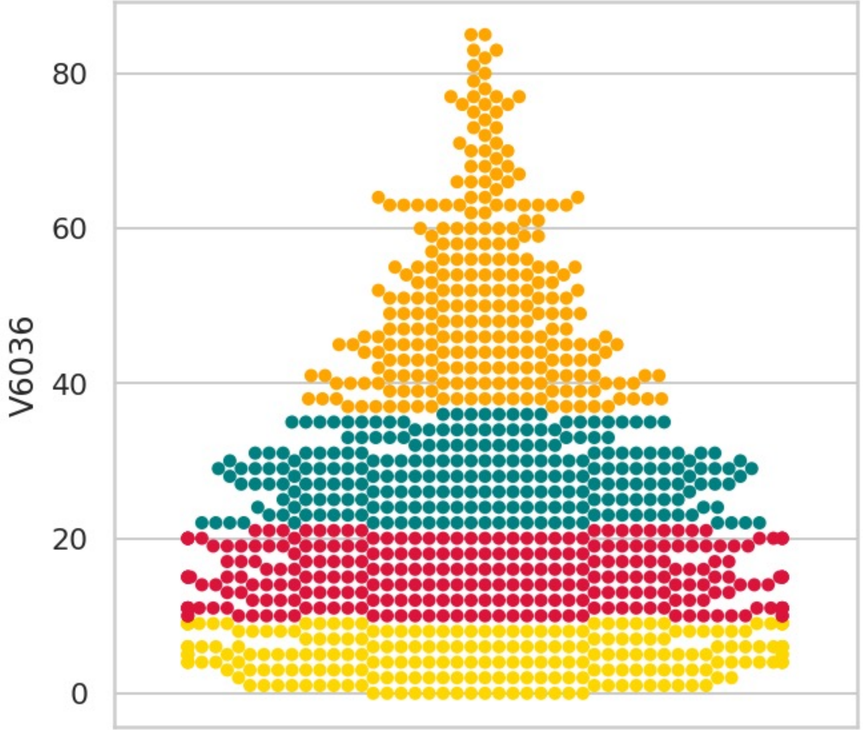
Então, fazemos uso da média entre os elementos nas posições 2 e 3

Q1 = (12+13)/2 = 12,5

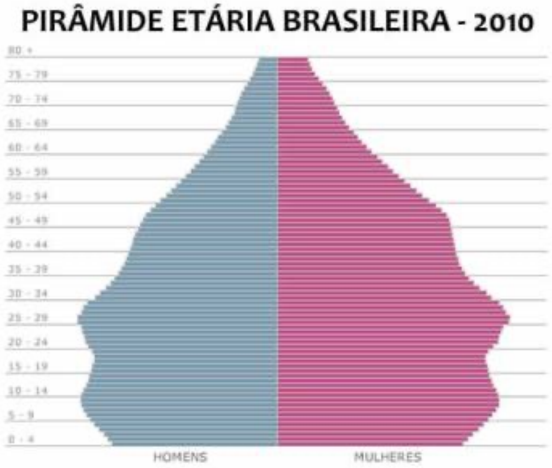
Medidas de Posição

Percentis e Quartis

- Exemplo: Censo IBGE 2010 – idade da população
- Fragmento de 1000 registros aleatórios selecionados



Swarm-plot

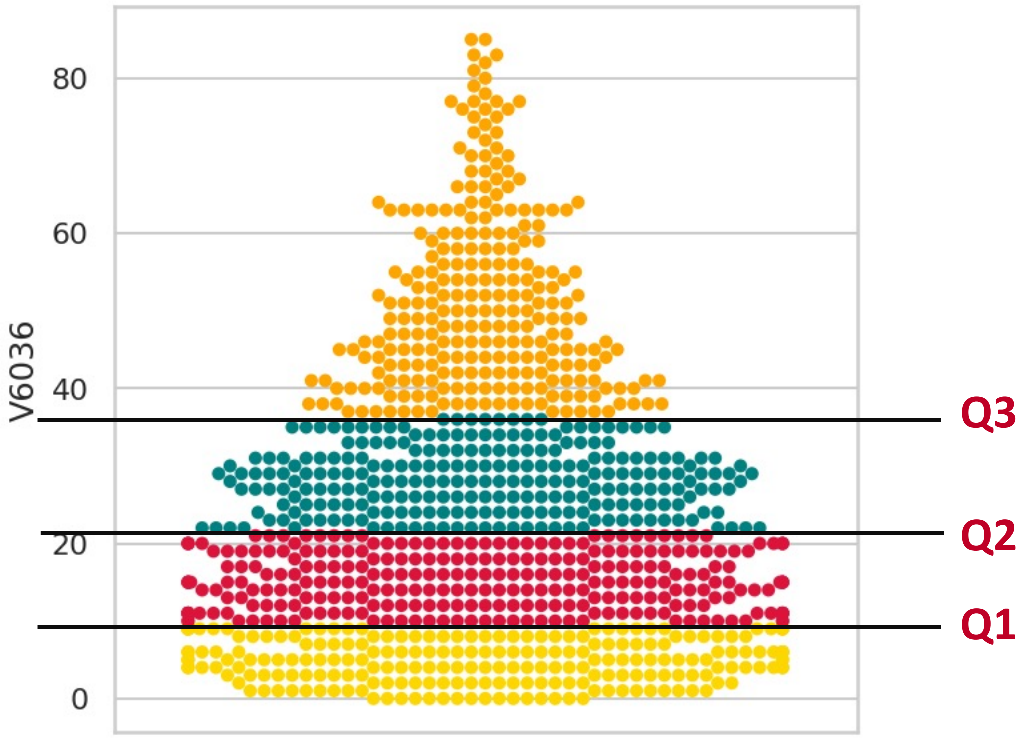


Insper

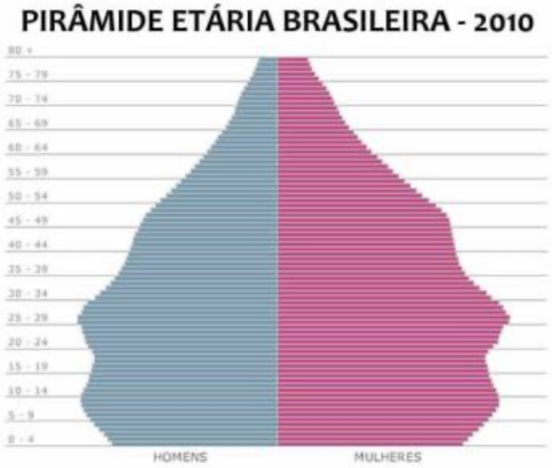
Medidas de Posição

Percentis e Quartis

- Exemplo: Censo IBGE 2010 – idade da população
- Fragmento de 1000 registros aleatórios selecionados



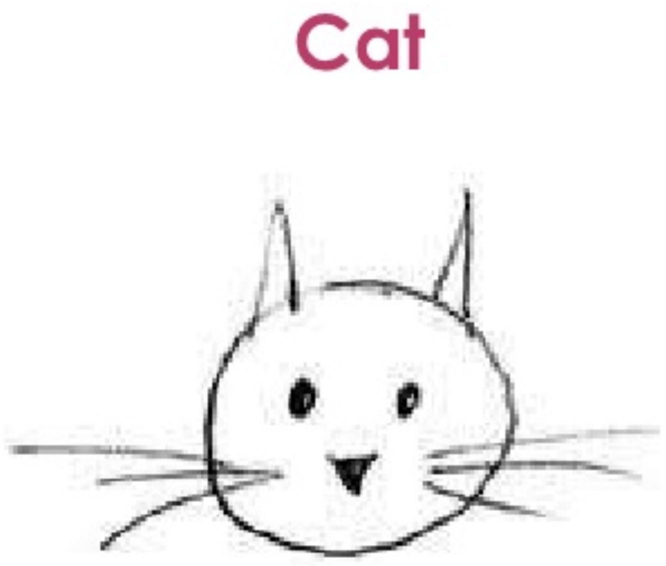
Swarm-plot

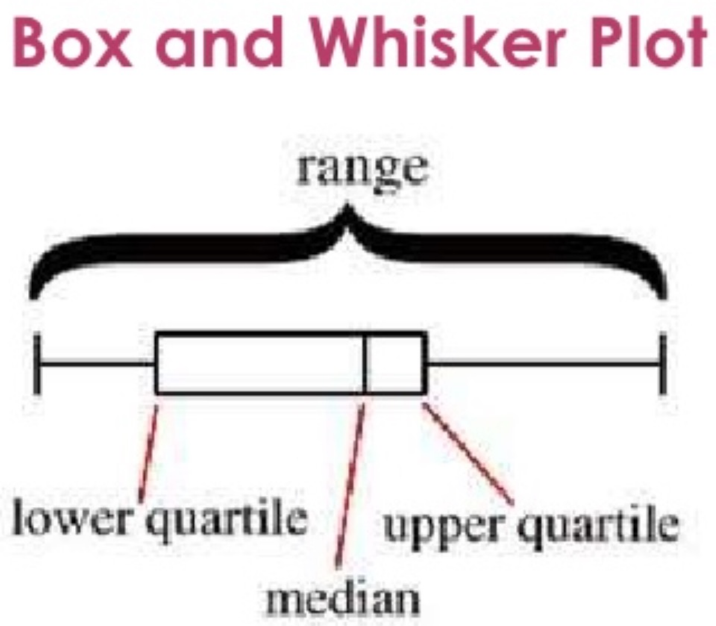


Visualizando Quartis

BoxPlot

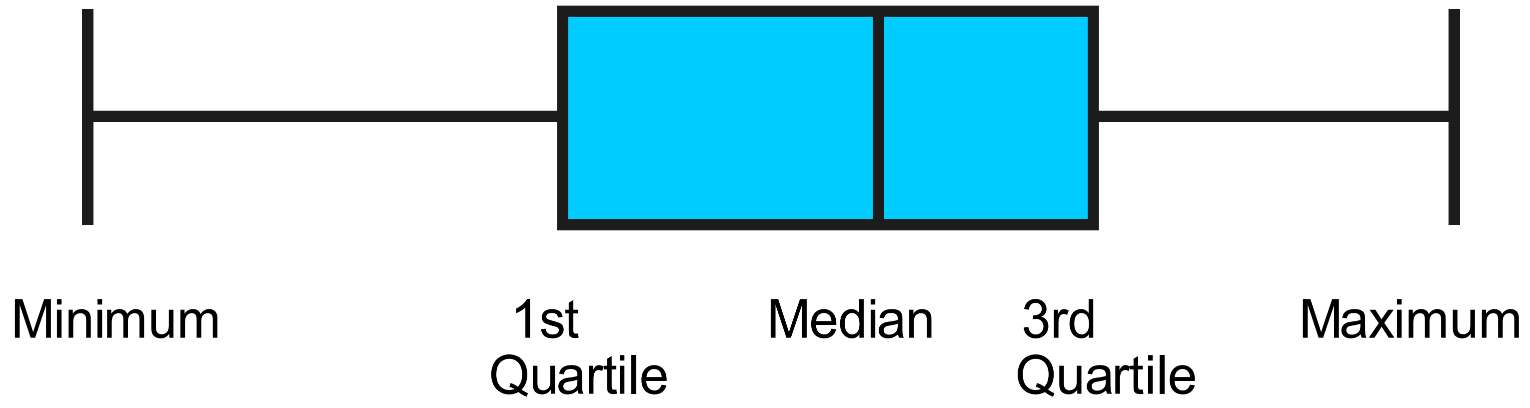
- Também conhecido como Box and Whisker Plot





Boxplot

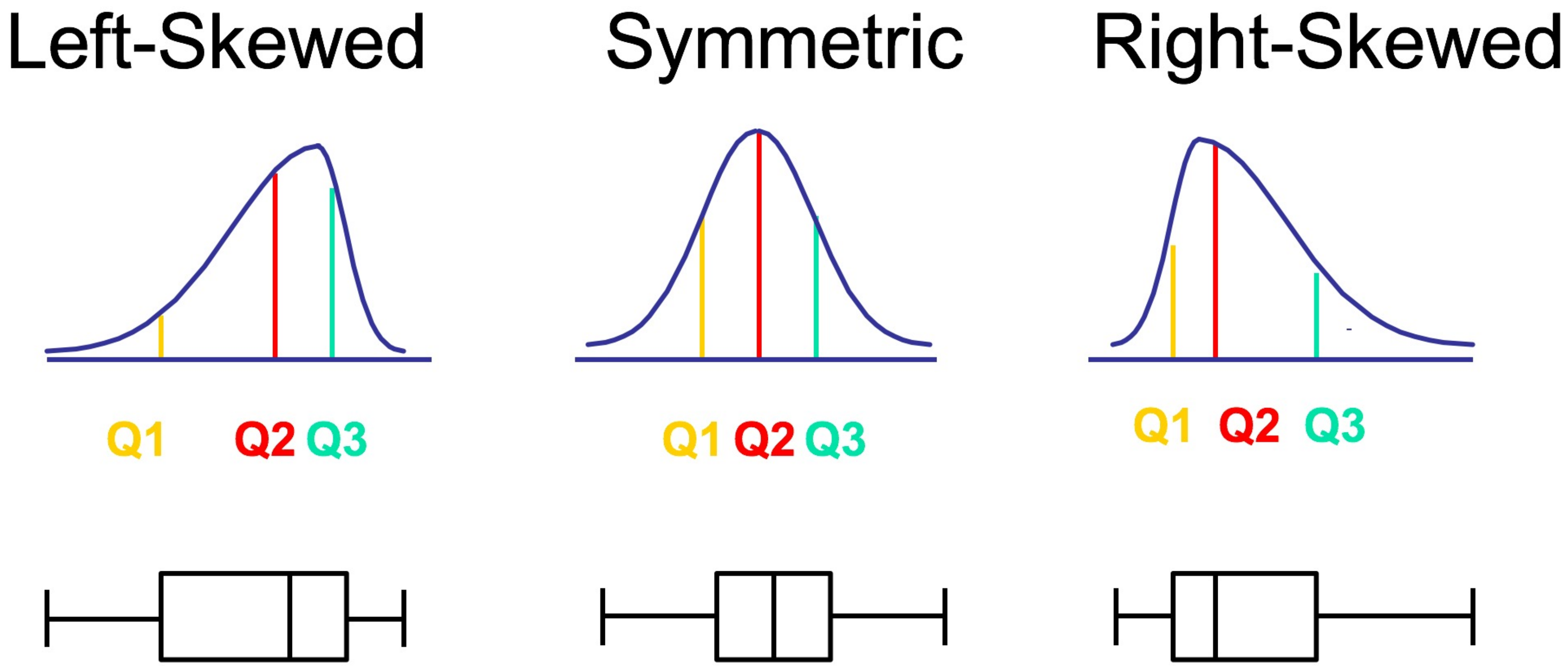
- Gráfico que apresenta os quartis de um conjunto de dados
- Em sua configuração clássica, pode ser representado da seguinte maneira



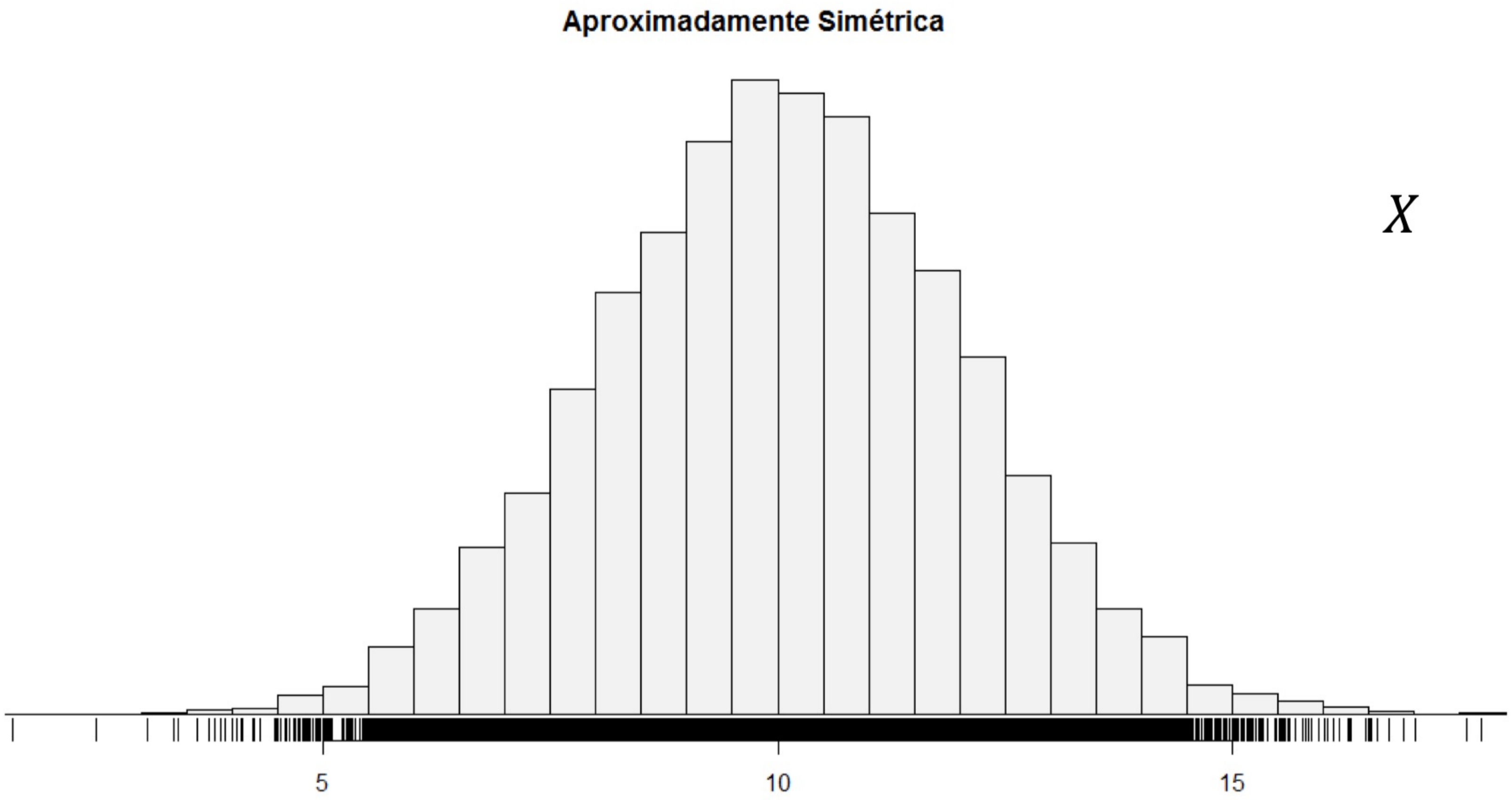
Obs: sem considerar outliers

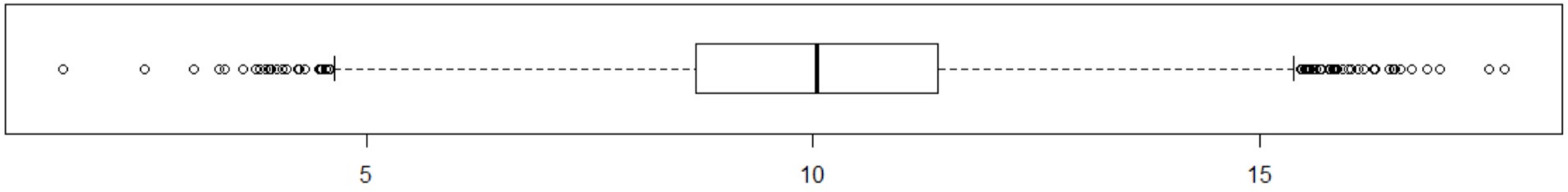
Boxplot

Boxplot e a forma da distribuição

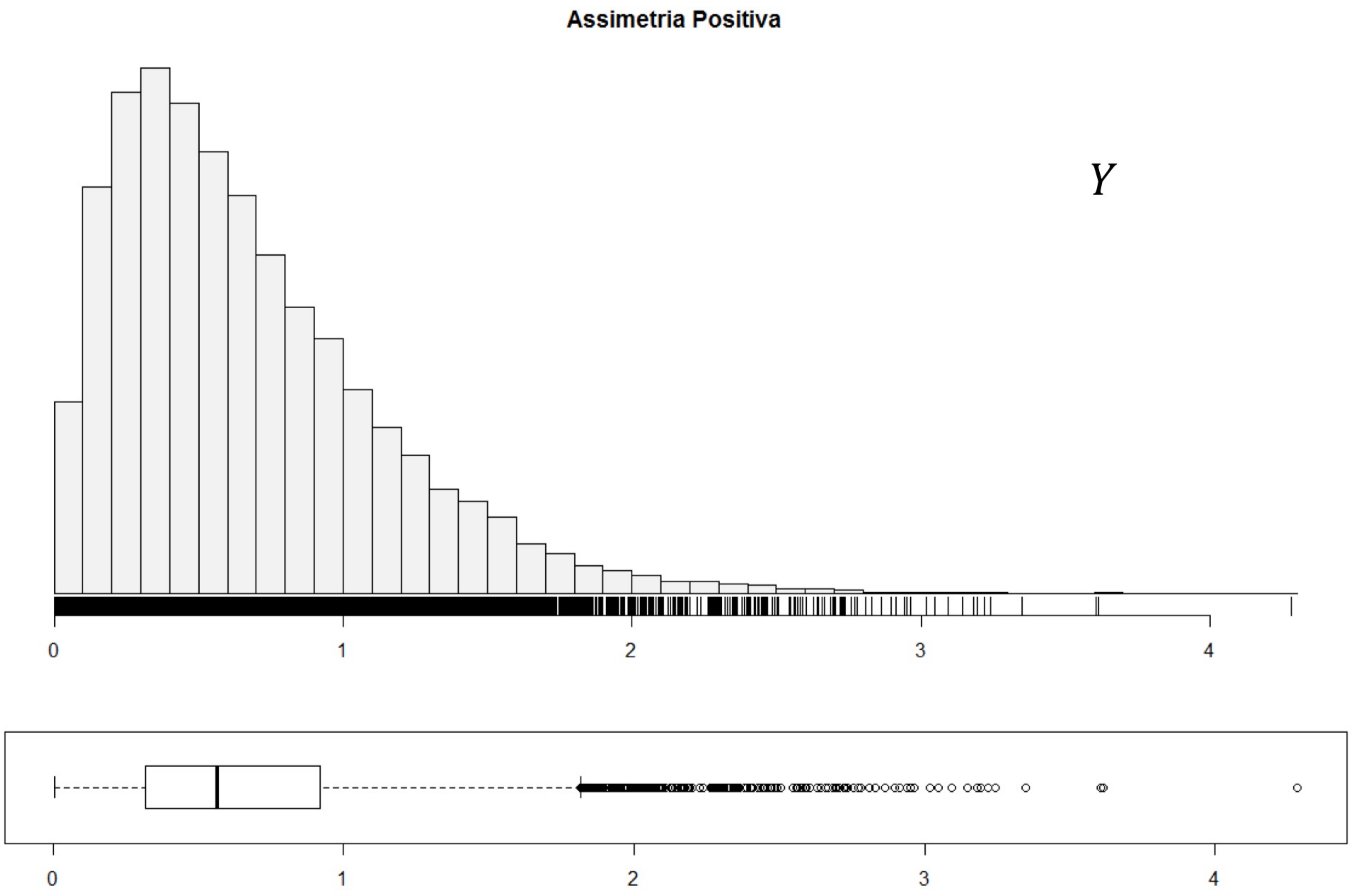


Histogramas e Boxplots



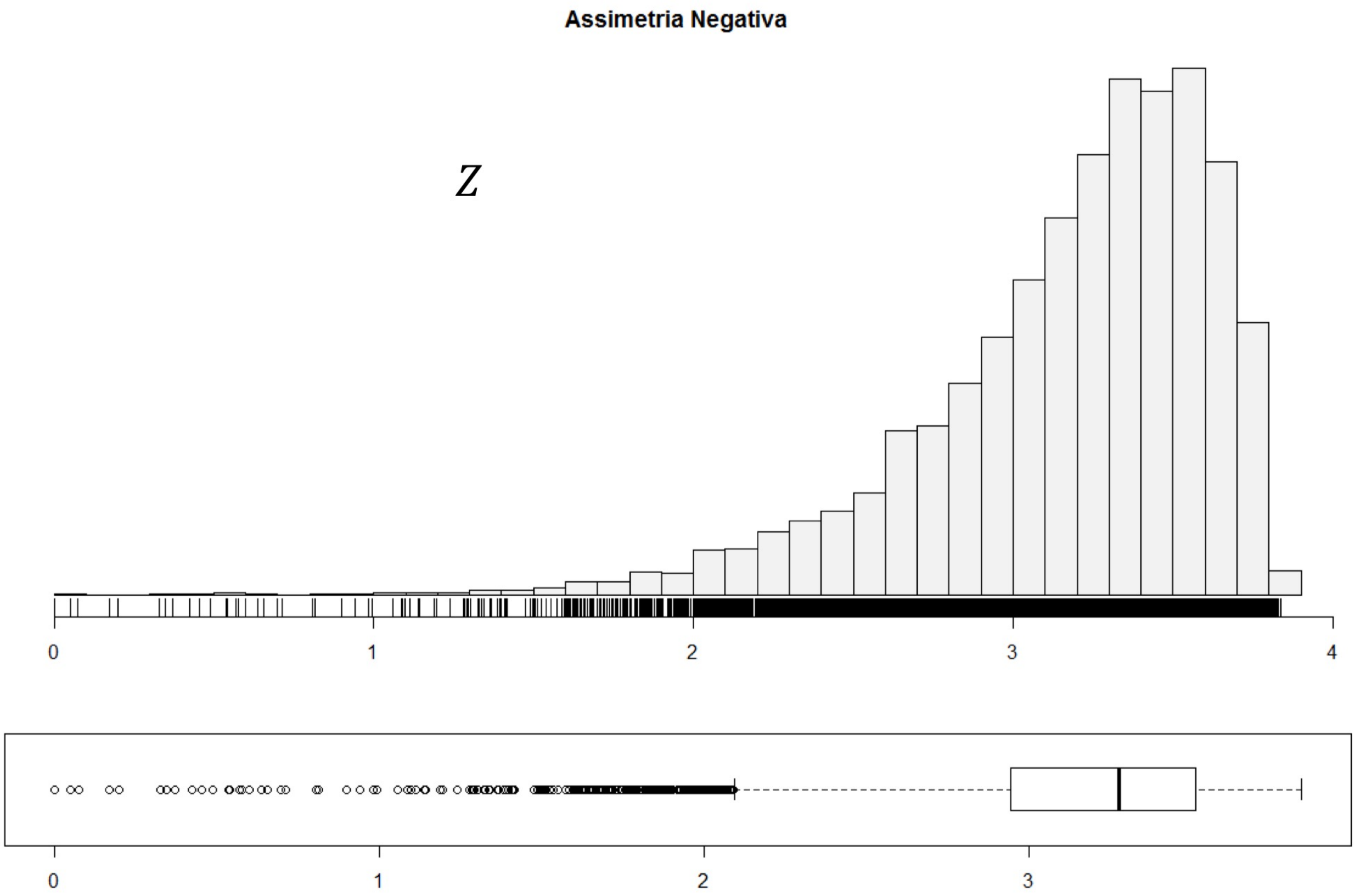


Histogramas e Boxplots



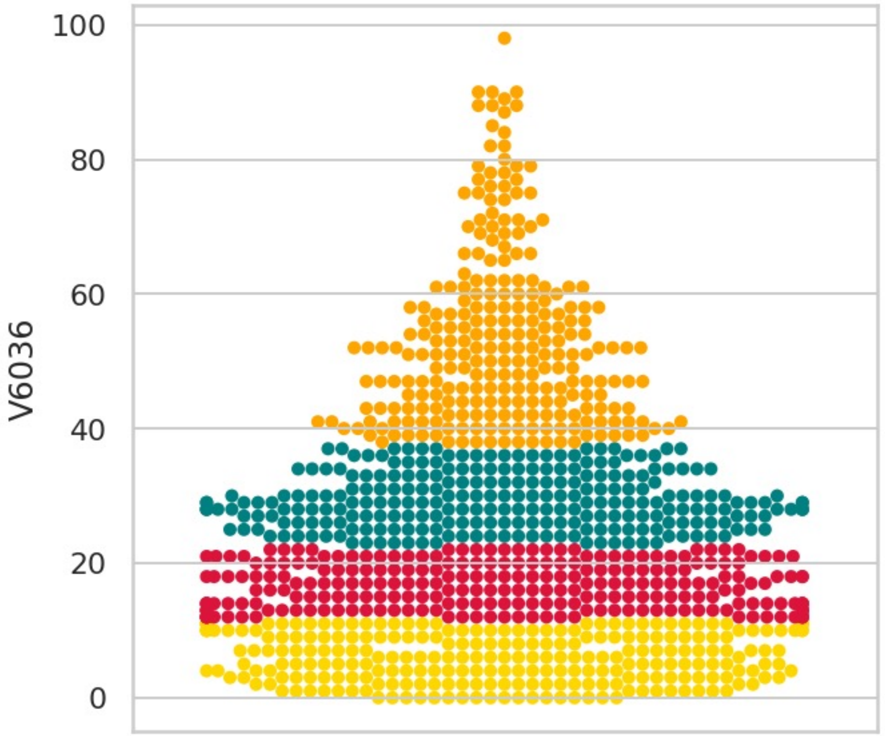
Insper

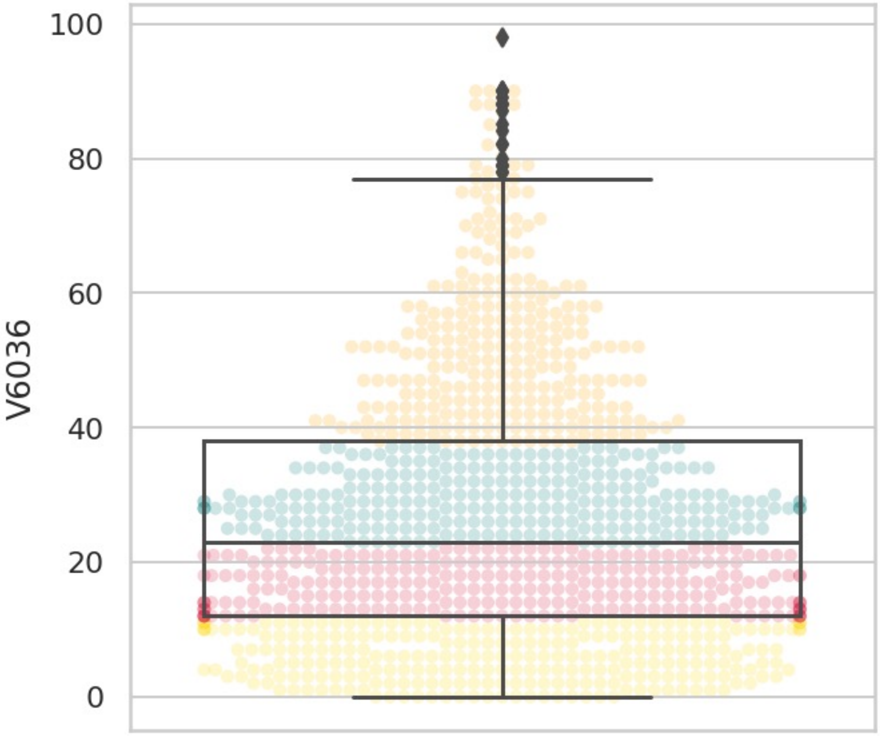
Histogramas e Boxplots



Insper

Boxplots

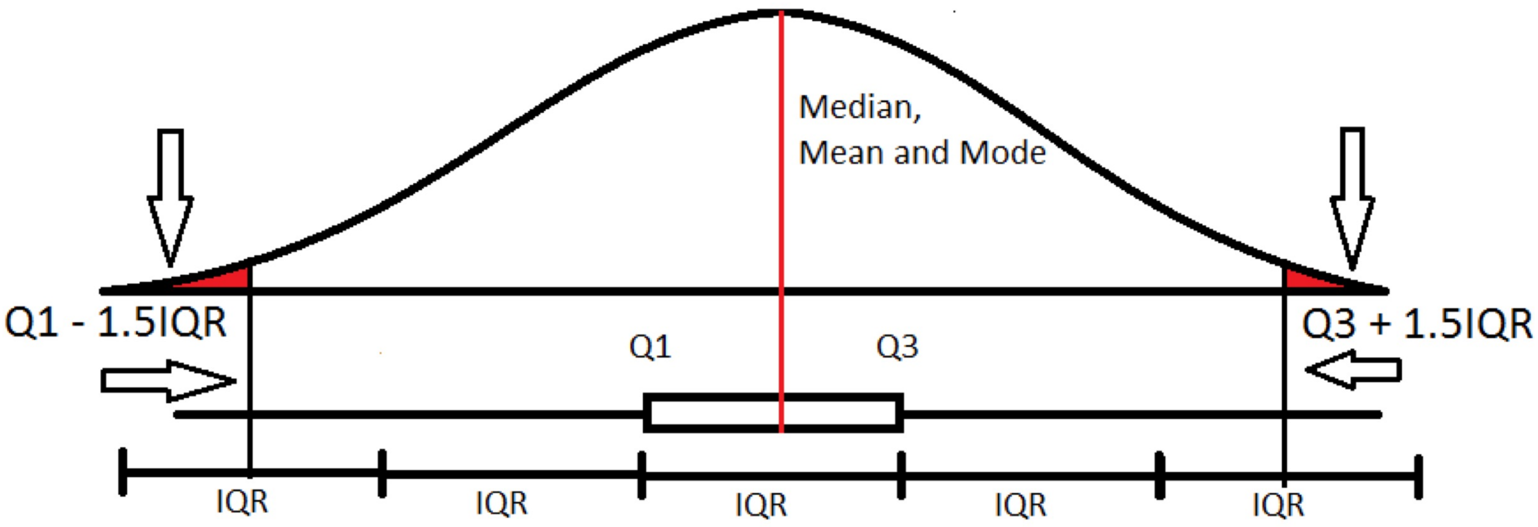




Boxplot

Outliers

- Uma forma de se obter os outliers em um conjunto de dados é usar o intervalo interquartil (IQR)
- Considera-se um outlier todos os valores:
  - Abaixo de Q1 – 1.5 x IQR
  - Acima de Q3 + 1.5 x IQR



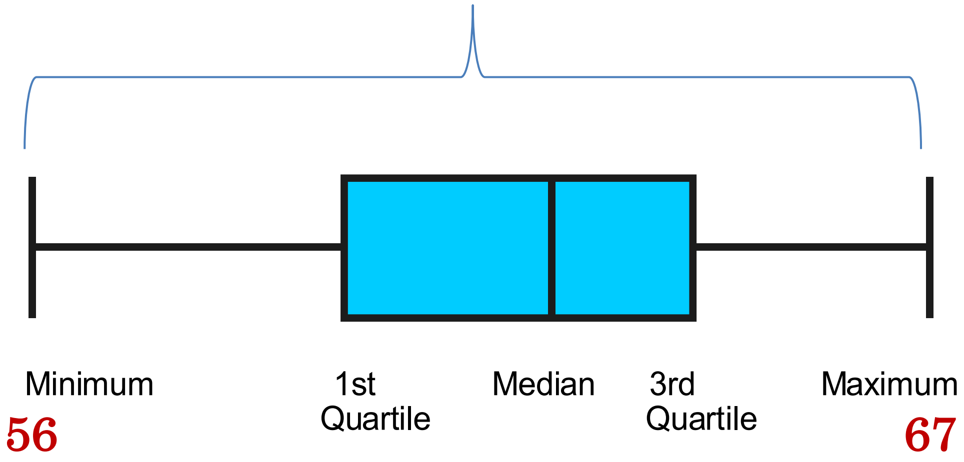
Medidas de dispersão

Amplitude

- A amplitude (range) de um conjunto de dados é a diferença entre o maior e o menor valor no conjunto

Bolsa	56	56	57	58	61	63	63	67	67	67

- A amplitude é 67 – 56 = 11



Insper

Medidas de dispersão

Desvio

- O desvio de uma entrada x em um conjunto de dados é a diferença entre x e a média μ dos dados

$$\text{Desvio de } x = x - \mu$$

- Ao lado temos valores sucessivos de uma certa bolsa de valores em um determinado período

- O valor médio é de 305 / 5 = 61

Bolsa x	Desvio $x - \mu$
56	$56 - 61 = -5$
58	$58 - 61 = -3$
61	$61 - 61 = 0$
63	$63 - 61 = 2$
67	$67 - 61 = 6$
$\Sigma x = 305$	$\Sigma(x - \mu) = 0$

Insper

Medidas de Dispersão

Variância e Desvio Padrão

- Representam a essência do conceito de variabilidade

- Levam em consideração todos os resultados existentes na distribuição

- A variância é uma medida de variabilidade que indica o grau em que todos os valores de uma distribuição se desviam da média. O desvio padrão é a raiz quadrada da variância, expresso portanto na mesma unidade dos dados, facilitando a compreensão
  - Quanto maior, mais os valores se distanciam da média
  - Desvio padrão baixo → grupo é homogêneo
  - Desvio padrão alto → grupo é heterogêneo

Medidas de Dispersão

Variância e Desvio Padrão

Bolsa x	Desvio $x - \mu$	Quadrado $(x - \mu)^2$
56	– 5	25
58	– 3	9
61	0	0
63	2	4
67	6	36
$\Sigma x = 305$	$\Sigma(x - \mu) = 0$	$\Sigma(x - \mu)^2 = 74$

$$SS_2 = \Sigma(x - \mu)^2 = 74$$

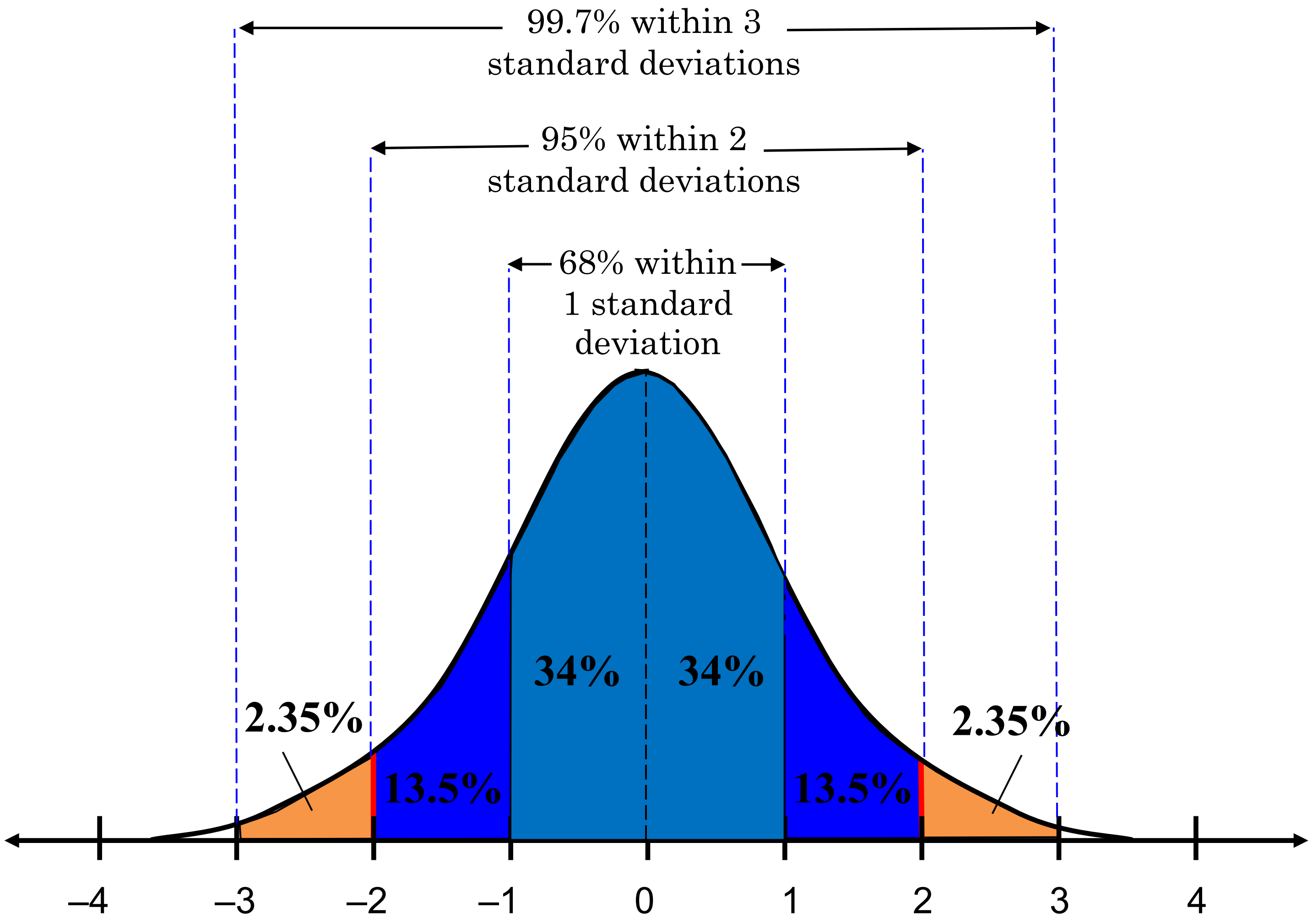
$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N} = \frac{74}{5} = 14.8$$

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} = \sqrt{14.8} \approx 3.85$$

$$\sigma \approx \$3.85$$

Empirical Rule

68 – 95 – 99.7%



Insper