

Análise Exploratória de Dados

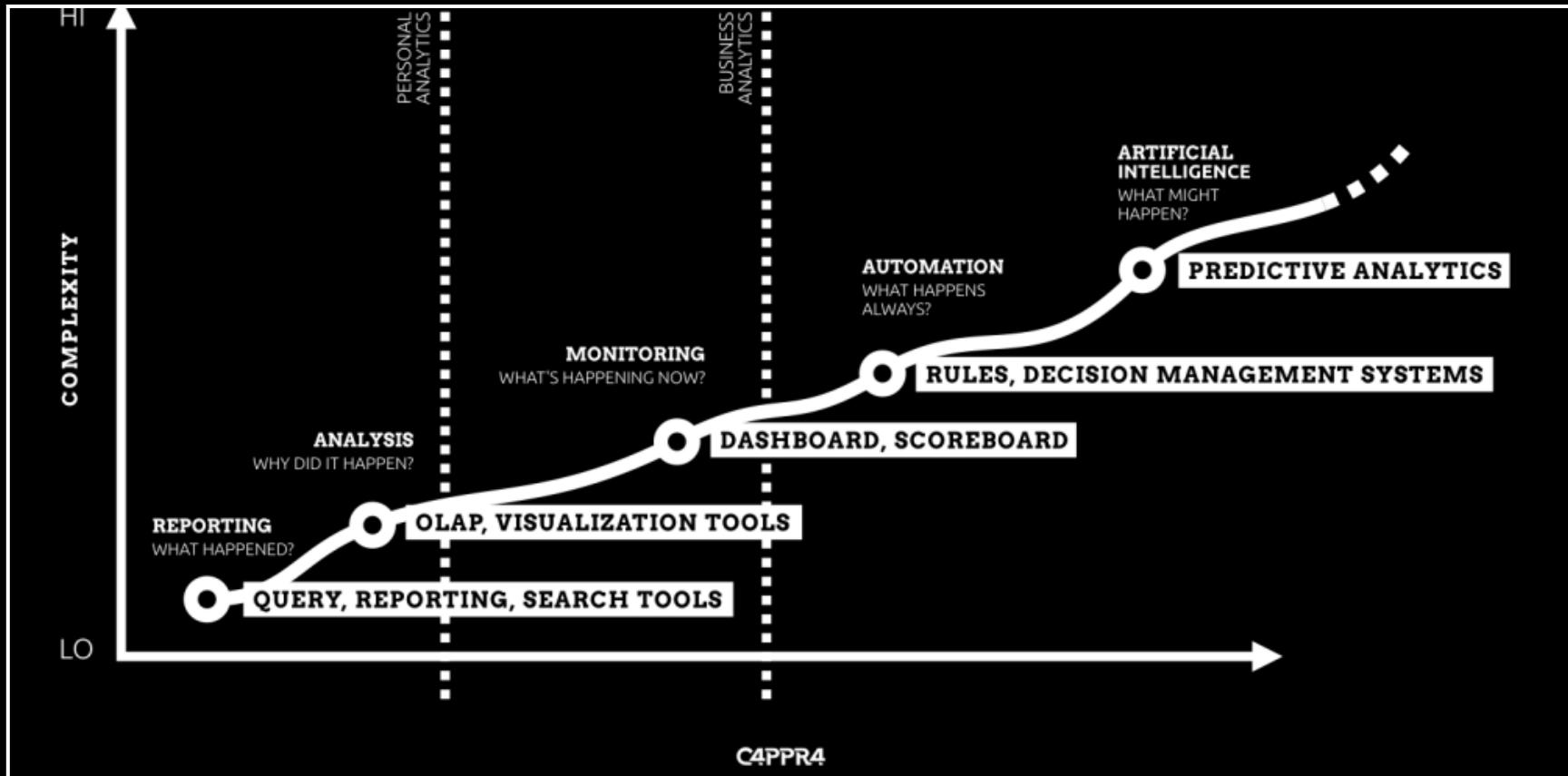
Danilo Augusto Cleto Souto

Apresentação

- Danilo Augusto Cleto Souto
 - Sistemas de Informação / Integração de Sistemas / Modelos Preditivos / Inteligência Artificial e Analytics
 - Celepar desde 2010
 - Linguagens
 - **R, Python, Java, C, C++, PHP**
 - Sistema de Gestão Hospitalar GSUS
 - Laboratório e Automação Laboratorial
 - Telefonia
 - Core Telefonia IP, Tarifação e SMS
 - GIE
 - Business Intelligence
 - Análise de Dados
 - Inteligência Artificial

*"I keep saying the sexy job in the next ten years will be **statisticians**. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?"*

2009, Google's Chief Economist, Hal Varian





MAY 6TH–12TH 2017

The Economist

Crunch time in France

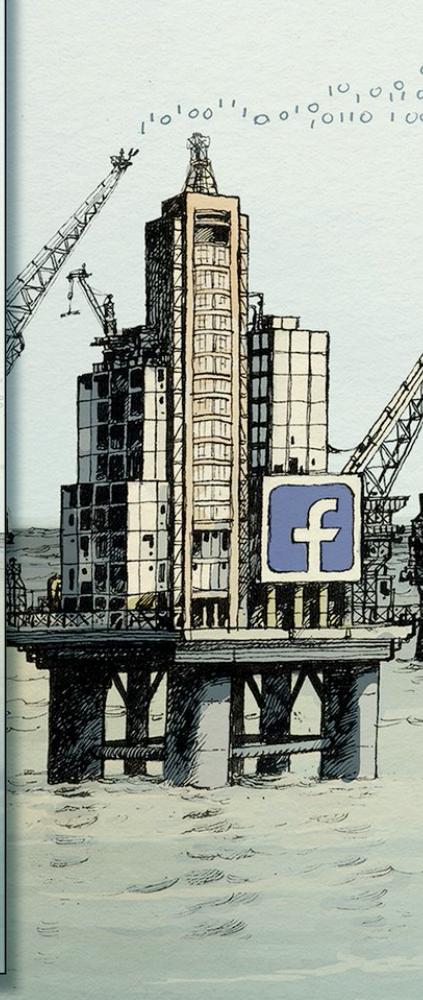
Ten years on: banking after the crisis

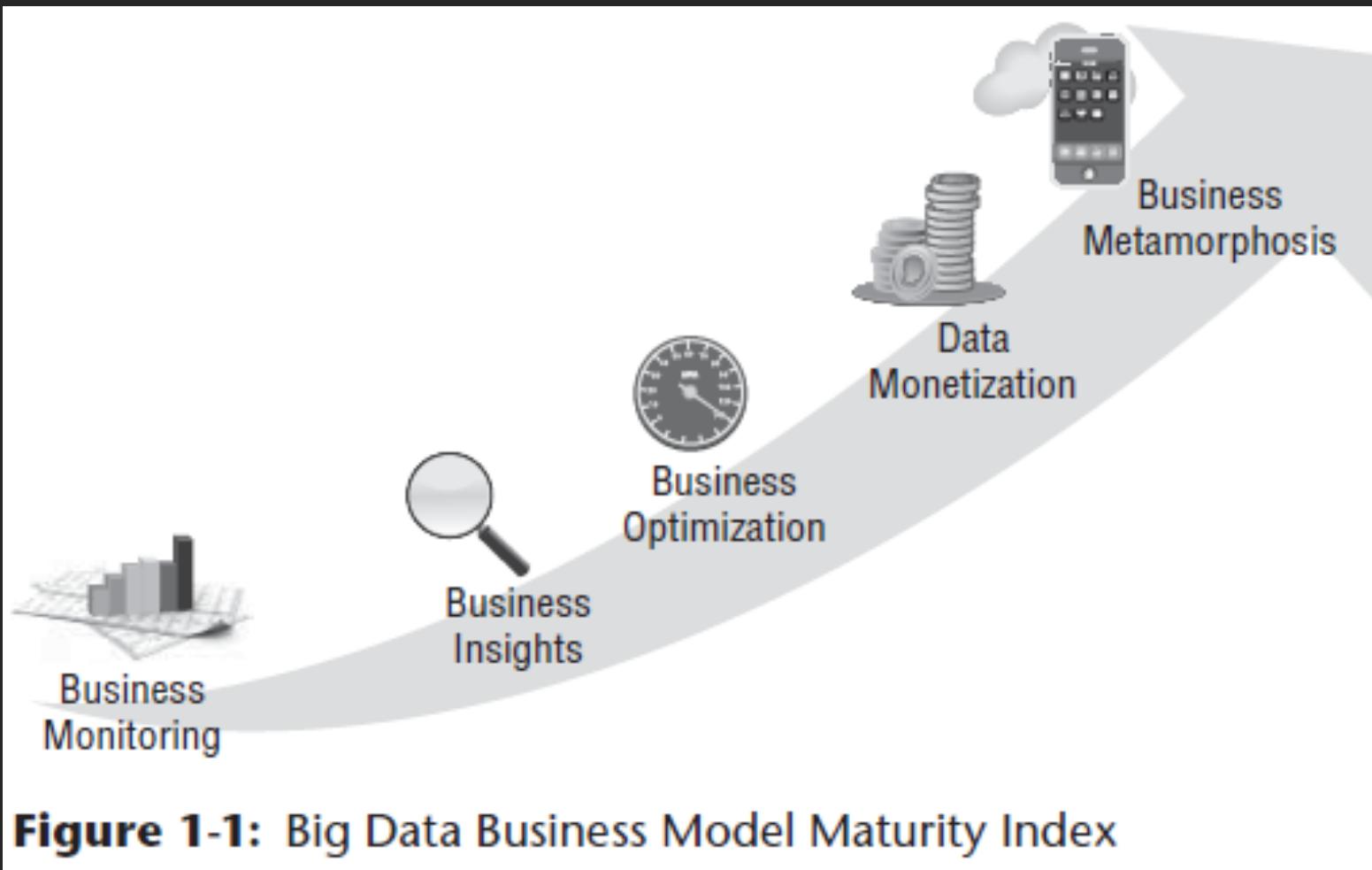
South Korea's unfinished revolution

Biology, but without the cells

The world's most valuable resource

Data and the new rules of competition





Learning

Sense and respond capabilities that use a mix of machine learning and AI methods to continually adapt and optimize to dynamic conditions

Deciding

Automated sense and respond capabilities tightly integrated into information governance and business processes

Advising

Model-based decision capabilities to support decision makers; SLAs forming; low business process impact

Orienting

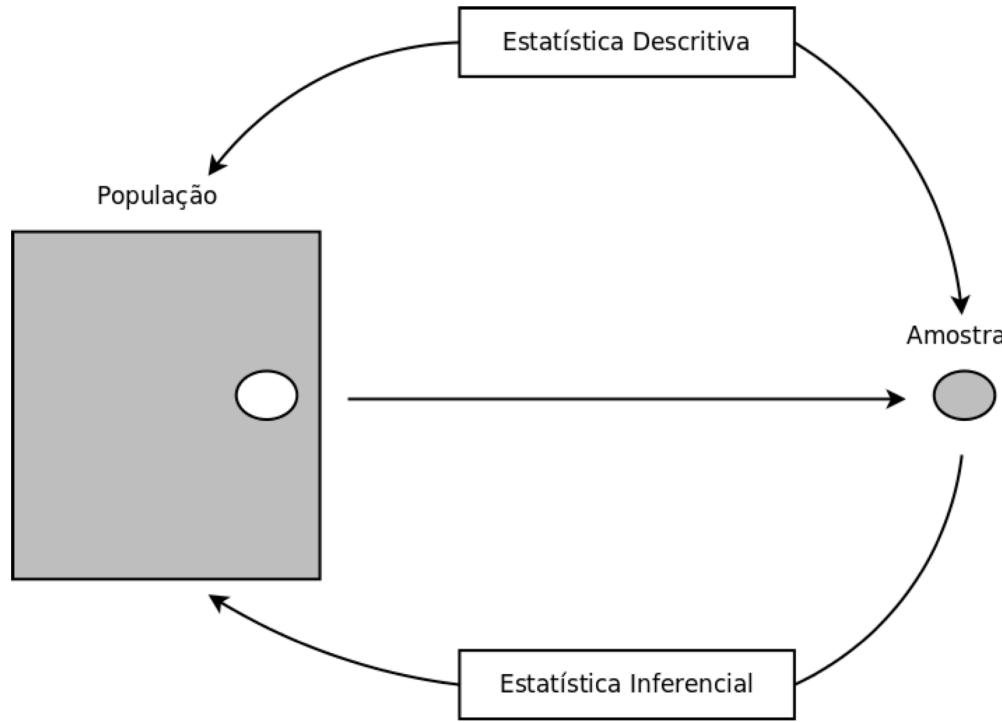
LOB-centric situational awareness; fragmented, starting analytics

Ingesting

Application-centric, ad hoc efforts; focus on data

Amostragem

Método para coletar dados de uma **pequena parte** de um grupo muito grande e “**aprender**” sobre esse **grupo muito maior**



População

- Conjunto de todos os elementos que apresentam uma ou mais características em comum;
- Fatores limitantes
 - População infinita
 - Custo
 - Tempo
 - Processos destrutivos
- **Parâmetro**
 - valor ou medida de uma característica populacional

Amostra

- Subconjunto de elementos extraídos de uma população
- Parâmetros populacionais desconhecidos
- Rápido
- Barato
- Amostragem é a técnica para extrair elementos e obter a amostra
- Estimativa
 - São valores estabelecidos para uma característica da amostra
- Subconjunto Representativo
- Aleatória

- População => Censo => Parâmetro
 - Os parâmetros são representados por letras gregas
- População => Amostra => Estimativa Estatística
 - Uma letra grega com acento circunflexo ou letra do alfabeto comum

Tipos de amostragem

- Probabilístico: todos os elementos possuem a mesma probabilidade (Aleatório)
 - Ex: Selecionar dez elementos por sorteio
 - Com Reposição: Elemento pode ser sorteado mais de uma vez
 - Sem reposição: o Elemento é amostrado somente uma vez
 - Sistemática: A cada N elementos 1 é retirado para amostragem (indústria)
 - Estratificada: População dividia em grupos (estratos) onde é realizado um sorteio dentro de cada grupo
- Não Probabilístico: Escolha deliberada de elementos
 - Conveniência: Elementos disponíveis
 - Julgamento: Escolhe intencionalmente os elementos
 - Não garante que a população seja representada

Exemplo

- Todos os Alunos da turma

Características altura

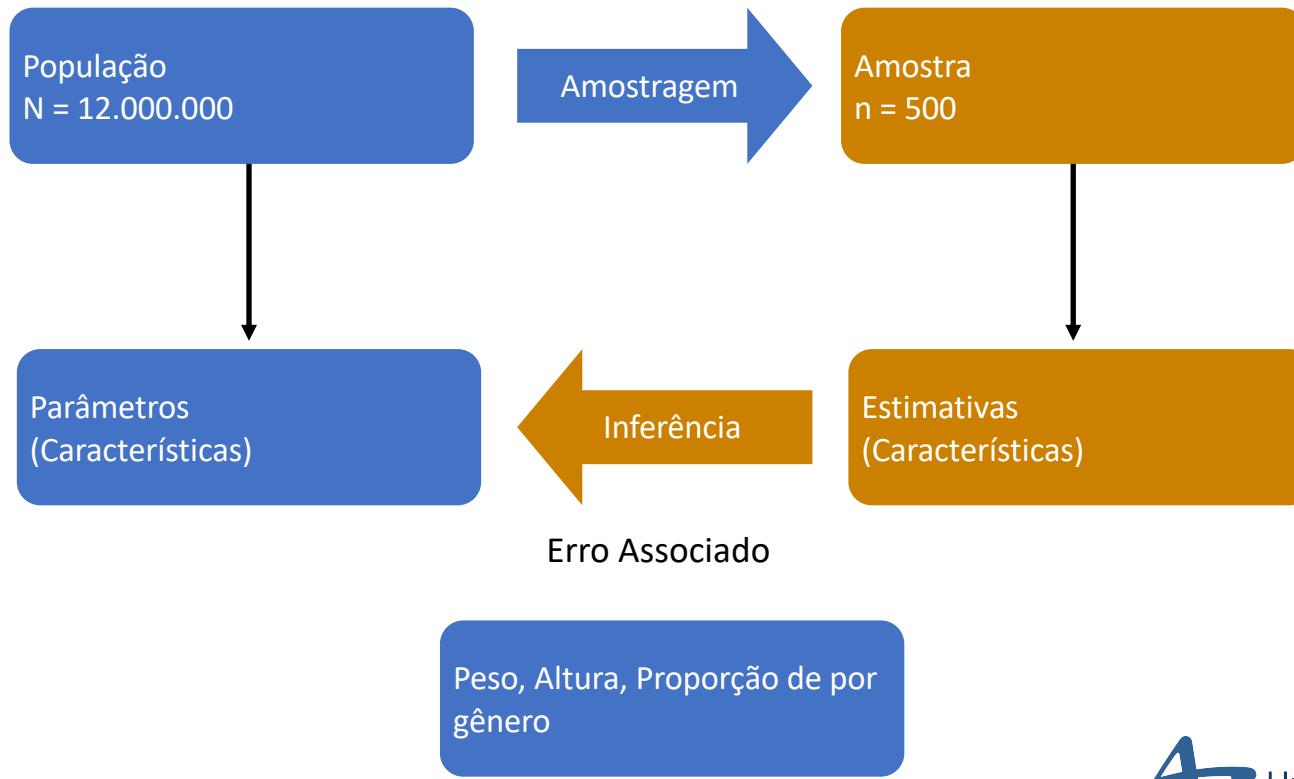
CENSO: [1.60, 1.74, 1.67, 1.71, 1.78, 1.71, 1.63, 1.67, 1.80, 1.73, 1.84, 1.63, 1.73, 1.90, 1.87, 1.87, 1.65, 1.64, 1.80, 1.70]

Média populacional: 1,733 <= Parâmetro

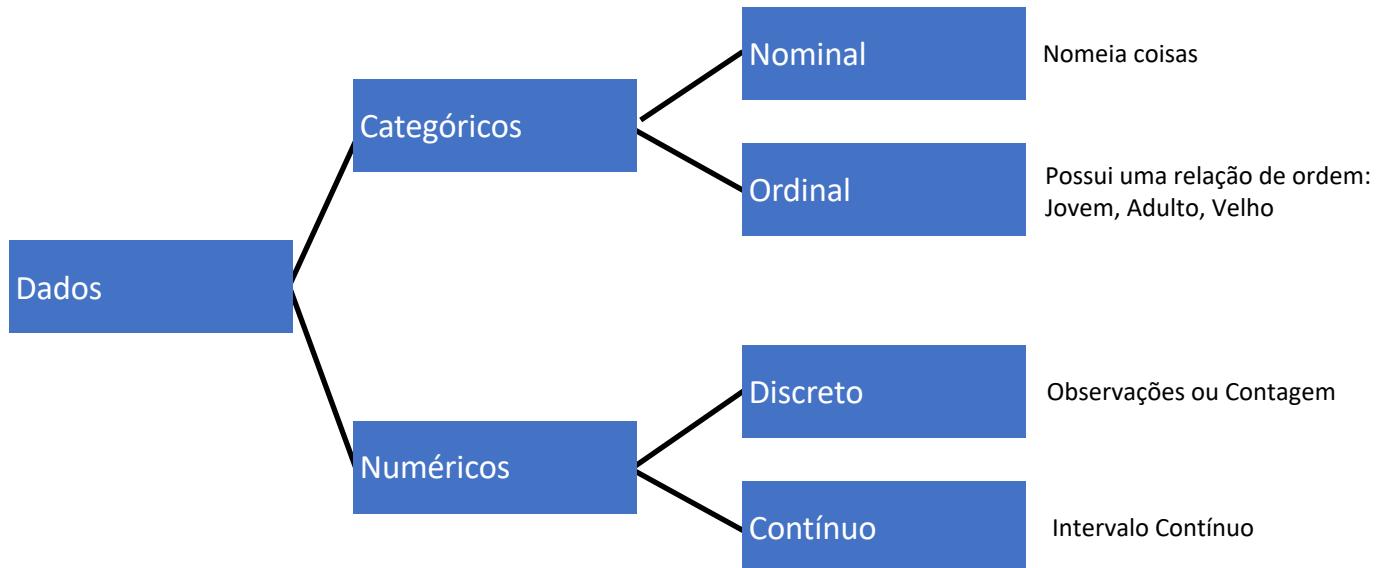
AMOSTRA [1.67, 1.67, 1.90, 1.87, 1.80, 1.60, 1.63]

Média amostral: 1,734 <= Estimativa

Exemplo



Natureza dos Dados



Organização

- Não pode haver perda de informação muito menos erro no processamento.
- A exibição deve ser de acordo com o tipo da variável a ser exibida
- Tabelas ou Gráficos podem ser escolhidos conforme o objetivos
- Matriz:
 - Linha: corresponde a cada elemento
 - Coluna: corresponde à variável

Matriz

Observação	Variável 1 (Idade)	Variável 2 (Altura)	Variável 3 (Peso)

- Variáveis Contínuas podem ser discretizadas se não perderem informação ou se o detalhamento for essencial para a análise:
 - Tempo: Mês, Semanas, Anos..
 - Quando não há precisão suficiente: Peso unidade em unidade.
- Quando uma informação estiver ausente deve-se utilizar um símbolo especial.
 - No R utiliza-se o **NA** - Not Available (Não disponível)
 - Nunca preencher com 0 (zero)

Análise Univariada

- Variável quanto ao tipo: Nominal, Ordinal ou Contínua
- Tabelas
- Gráficos
- Medidas para resumir a variável

Variável Categórica Nominal

- Tabelas de frequência
 - Absoluta
 - Relativa
- Gráfico de Barras ou Pizza

	freq	p
4 cilindros	11	0.34375
6 cilindros	7	0.21875
8 cilindros	14	0.43750

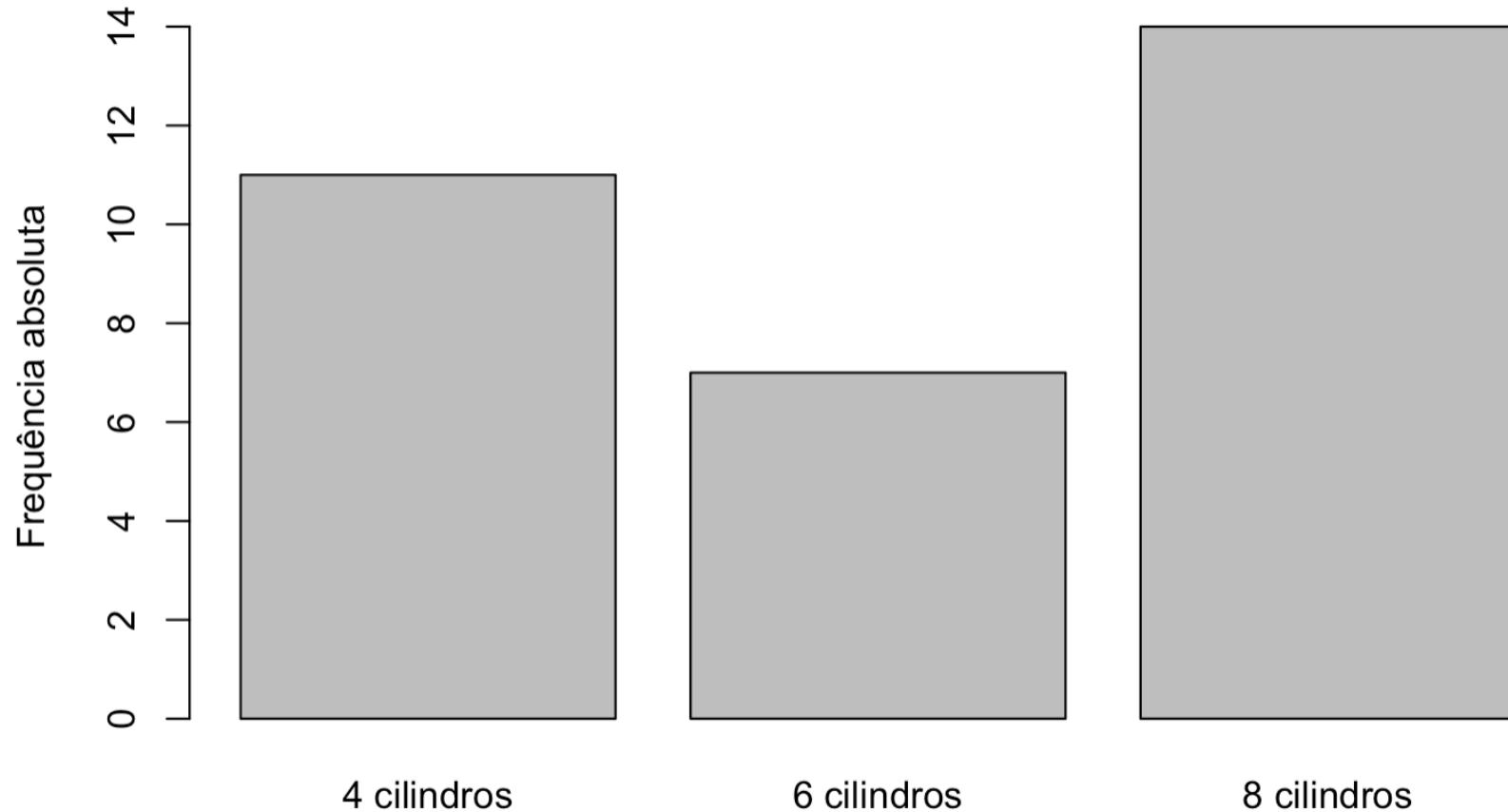
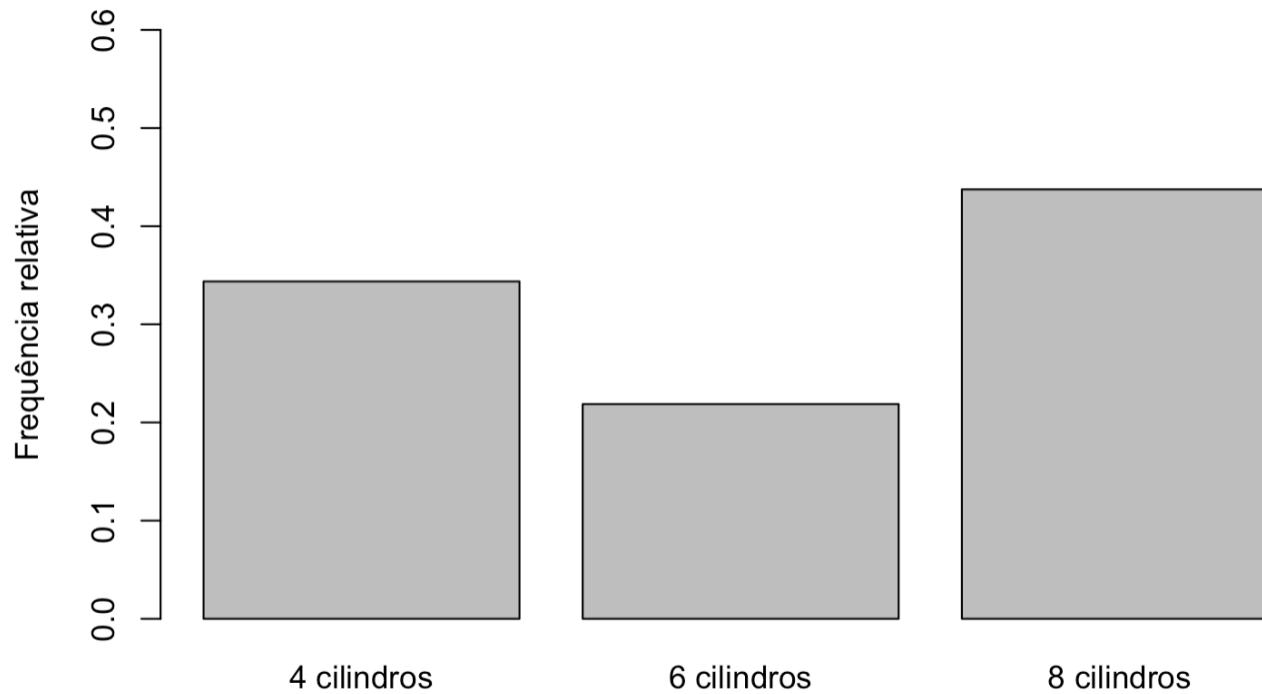
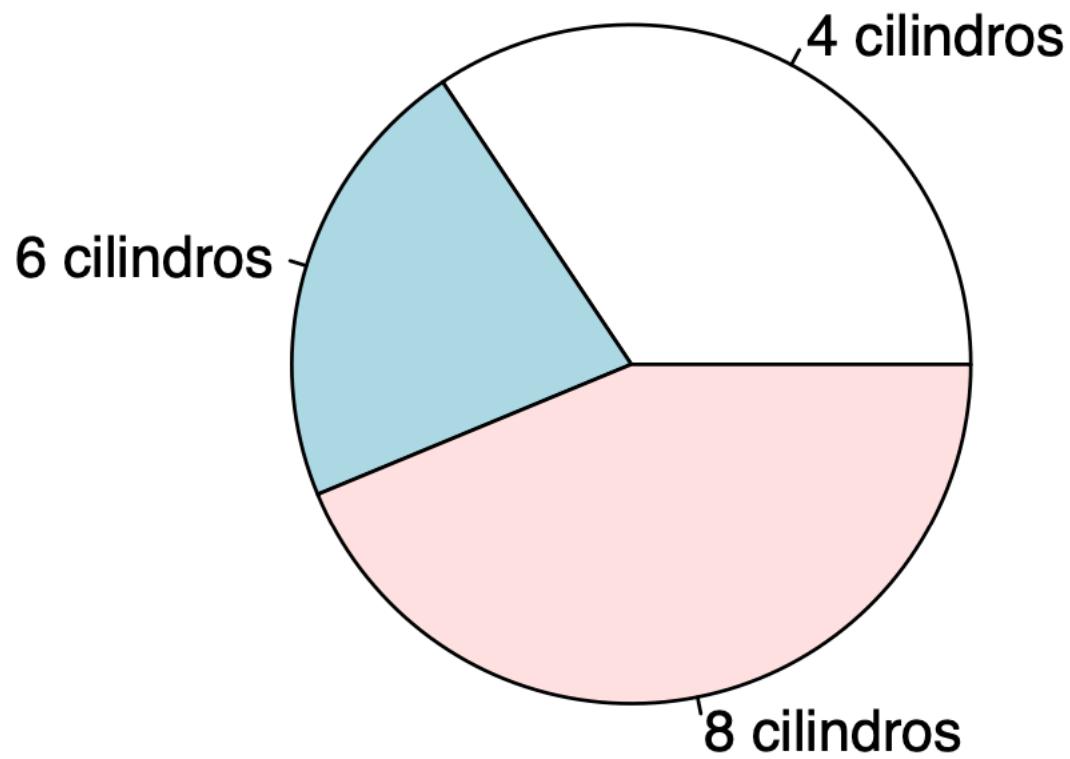


Gráfico de barras frequênci a relativa



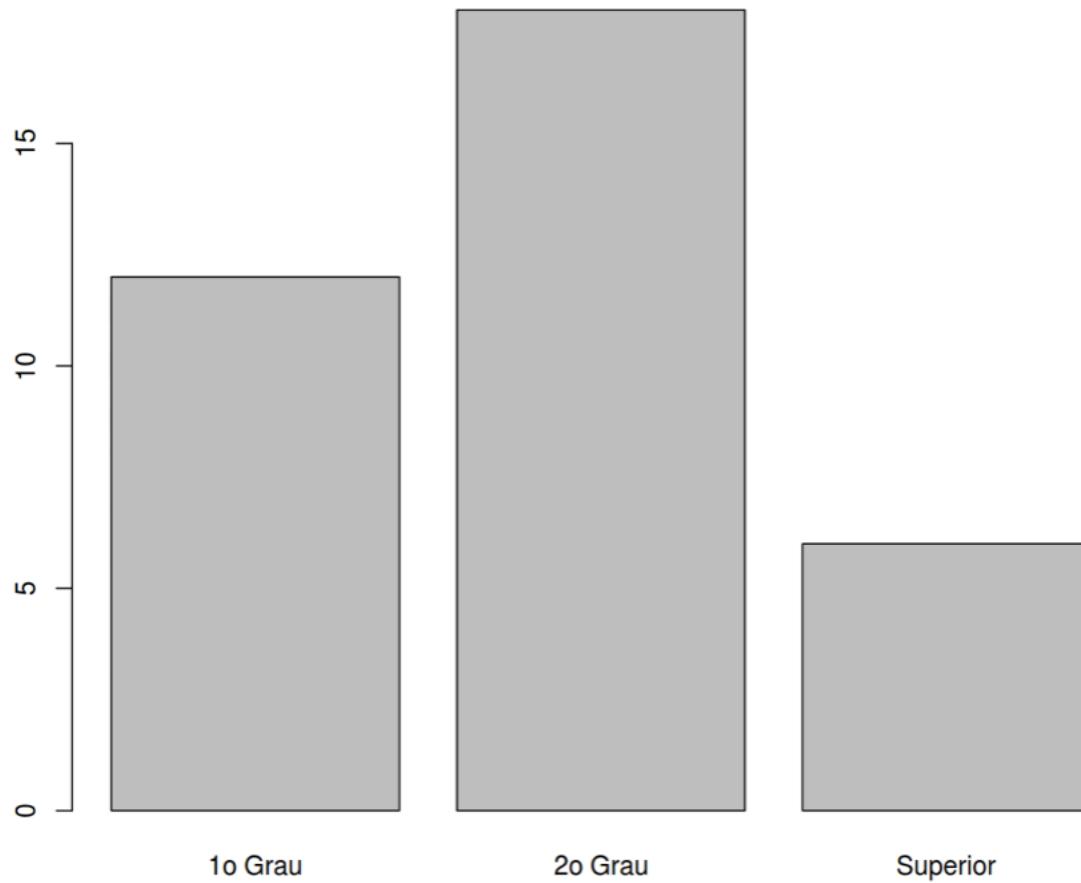


Variável Categórica Ordinal

- Tabelas de frequência
 - Absoluta
 - Relativa
- Gráficos de barras
- A ordem tem importância, então deve-se dispor a ordem na ordem natural das categorias.
- Não se utiliza gráfico de pizza por não representarem a ordem.

Tabela de frequências

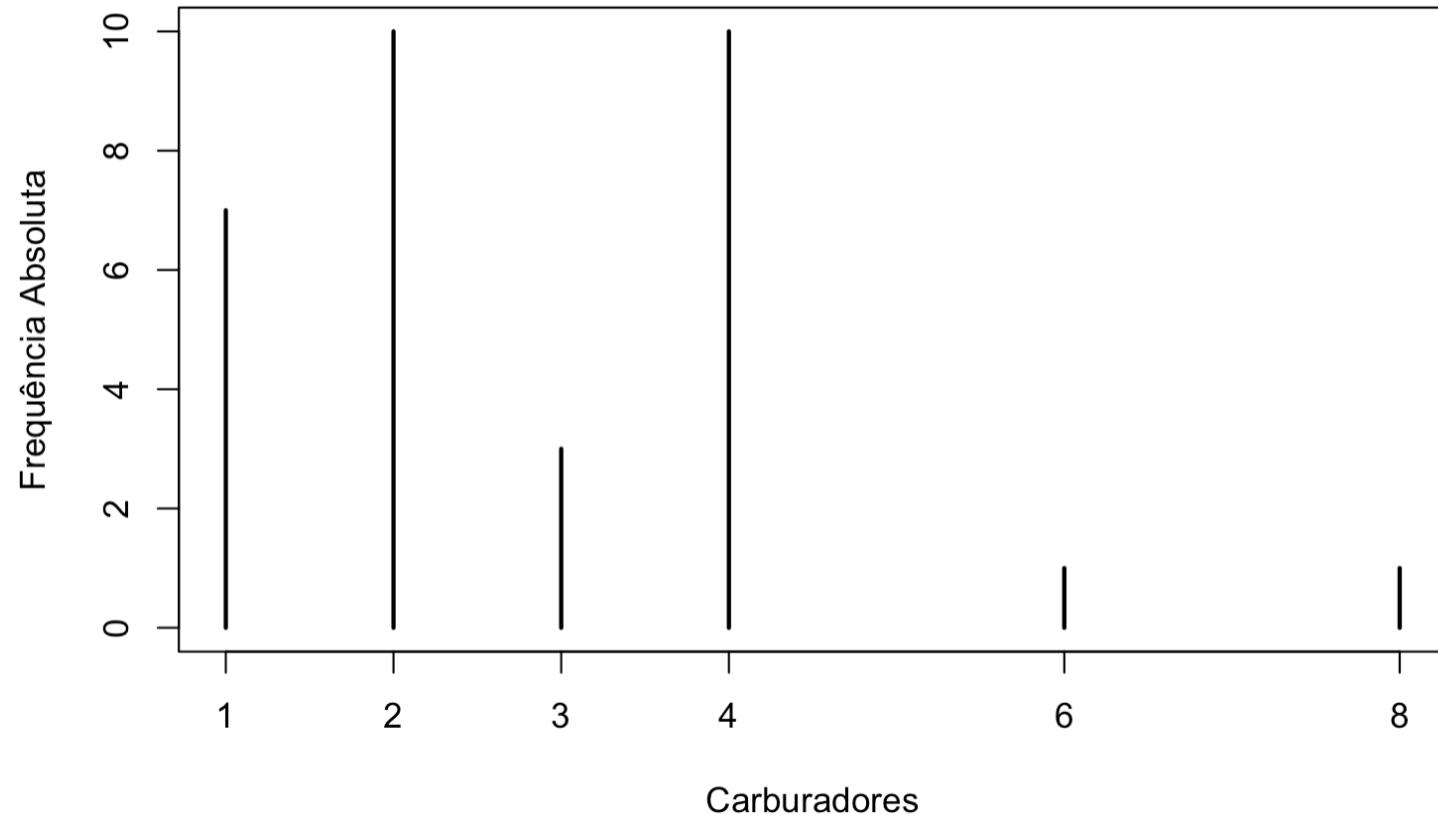
	f	F	r	R
4 cilindros	11	11	0.34375	0.34375
6 cilindros	7	18	0.21875	0.56250
8 cilindros	14	32	0.43750	1.00000

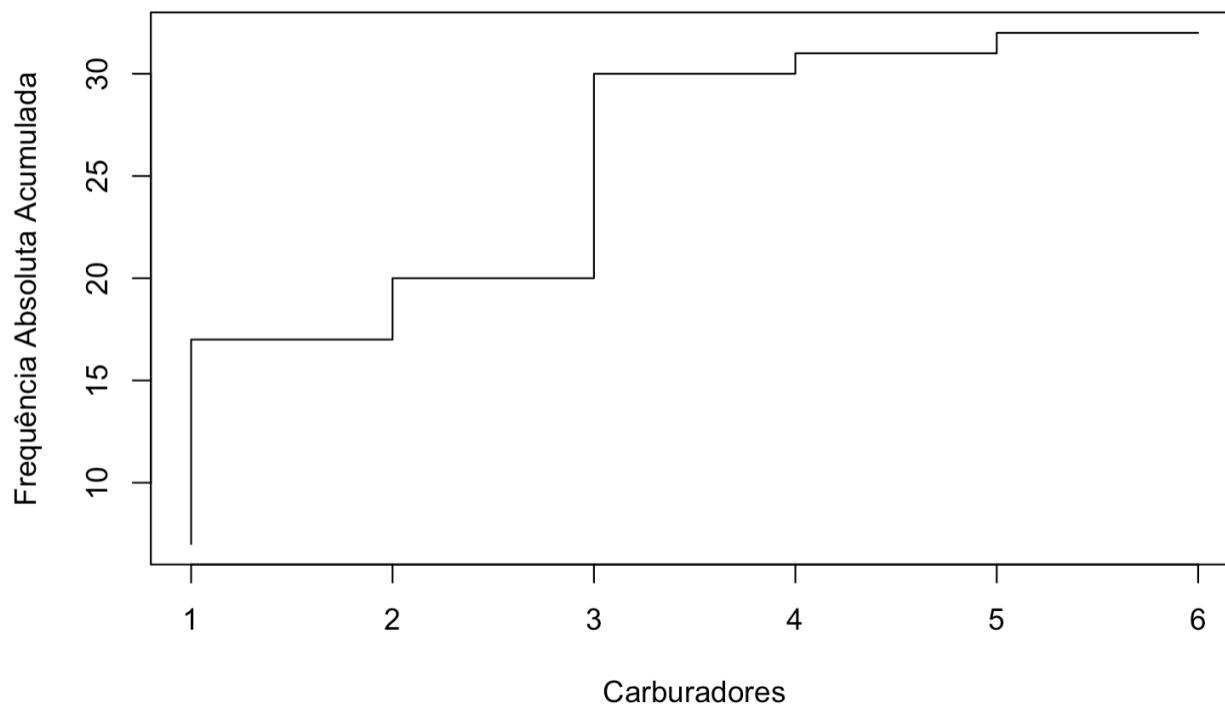


Variável Numérica Discreta

- Tabelas de frequências (poucos valores)
 - Absolutas
 - Relativas
 - Absoluta Acumulada
 - Relativa Acumulada

	f	p	F	R
1	7	0.21875	7	0.21875
2	10	0.31250	17	0.53125
3	3	0.09375	20	0.62500
4	10	0.31250	30	0.93750
6	1	0.03125	31	0.96875
8	1	0.03125	32	1.00000
Sum	32	1.00000	NA	NA





Variável Numérica Contínua

- Tabelas de frequência de classes

- Intervalo de classe

$$h = \frac{AT}{k}$$

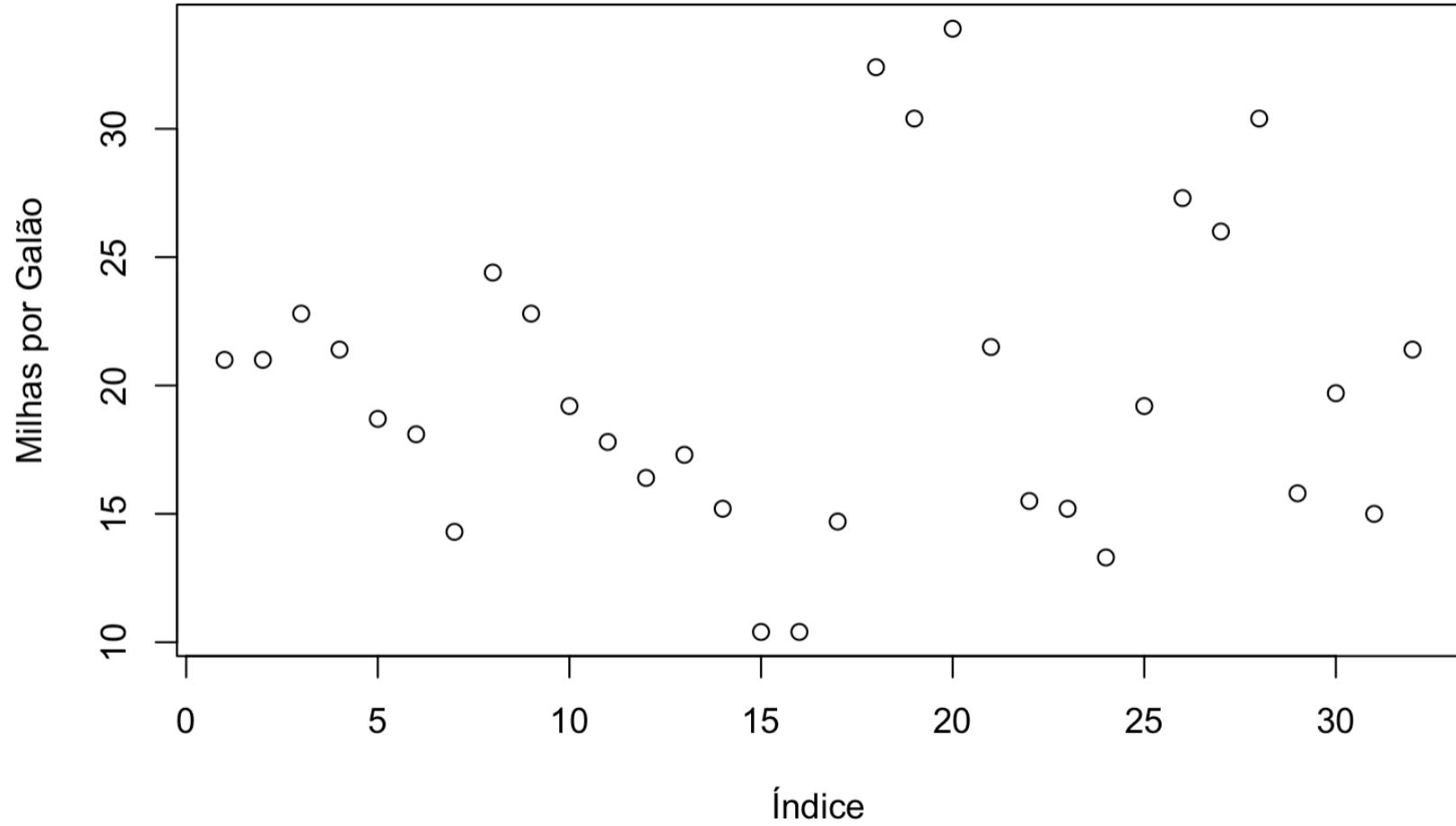
- Gráfico de dispersão
 - Variável x índice

onde $AT = \max x - \min x$ é a amplitude total dos dados, e $k = \sqrt{n}$ é

- Histograma
 - Frequência
 - Densidade

	Classe	Notação	Denominação	Resultado
	$[a,b)$	$a \vdash b$	Fechado em a, aberto em b	Inclui a, não inclui b
	$(a,b]$	$a \dashv b$	Aberto em a, fechado em b	Não inclui a, inclui b

	f	r	F	R
[10,14)	3	0.09375	3	0.09375
[14,18)	10	0.31250	13	0.40625
[18,22)	10	0.31250	23	0.71875
[22,26)	3	0.09375	26	0.81250
[26,30)	2	0.06250	28	0.87500
[30,34)	4	0.12500	32	1.00000
Sum	32	1.00000	NA	NA



Análise Monovariada

- Emprego de modelos quantitativos para descrição de amostras de dados relativos a um **único atributo**
- Perguntas que a análise monovariada pode responder:
Qual é o comportamento de um determinado fenômeno?
- O diretor de um hospital precisa determinar qual é a quantidade de atendimentos realizados no ambulatório, de maneira a avaliar a capacidade de atendimento instalada
- O gerente de qualidade de uma fábrica de artefatos plásticos deseja verificar se a resistência plástica dos produtos é estável (está sob controle)
- O coordenador de um curso deseja verificar se o desempenho de uma determinada turma está dentro das metas esperadas

Medidas de Centralidade

- Moda
- Mediana
- Média
- Resumem os valores centrais da amostra

Moda

- É o valor que mais se repete no conjunto de dados
- Contagem simples de valores
- Se o Conjunto possuir mais de uma moda é chamado bimodal
- Mais de duas modas, é chamado multi modal.

- Resistente a valores muito distoantes
- Pode ser utilizada para dados qualitativos

Mediana

- Valor central que divide a amostra em dois conjuntos de mesmo tamanho

4 elementos

4 elementos

i	x
1	5
2	10
3	16
4	21
5	25
6	33
7	38
8	47
9	58

Atenção: A mediana exige que os valores observados sejam ordenados.

Mediana = 25

Mediana

- Mediana para uma lista com número par de elementos utiliza-se a média dos valores centrais
- É resistente a valores altos

<i>i</i>	x
1	5
2	10
3	16
4	21
5	25
6	33
7	38
8	47
9	58
10	60

4 elementos

Elementos centrais

4 elementos

Atenção: A mediana exige que os valores observados sejam ordenados.

$Mediana = \frac{25 + 33}{2} = 29$

Média Aritmética

- Valor de Tendência central
- Uma medida consistente
- Sensível aos valores extremos da amostra

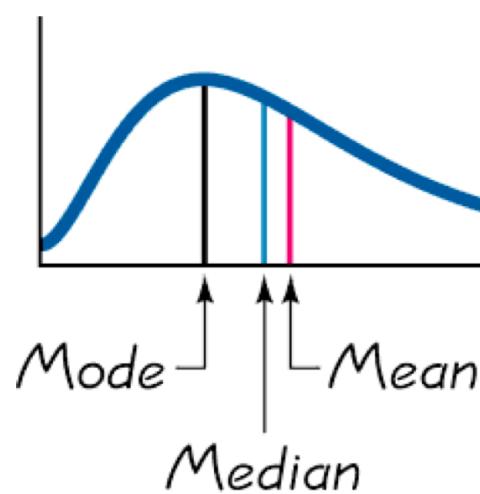
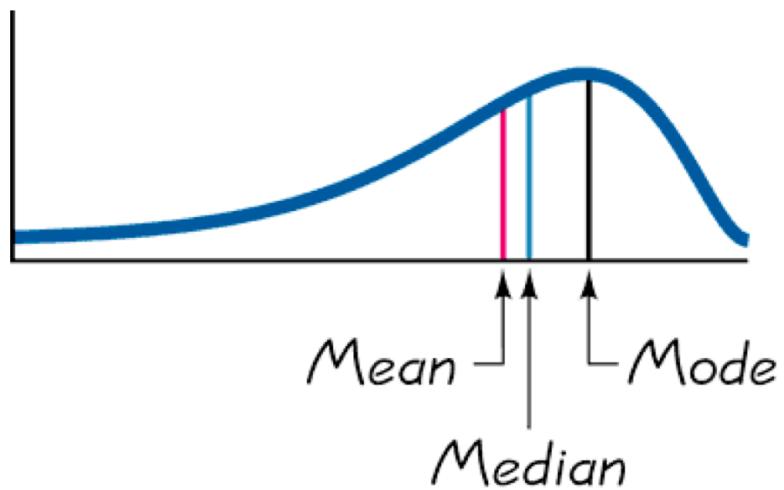
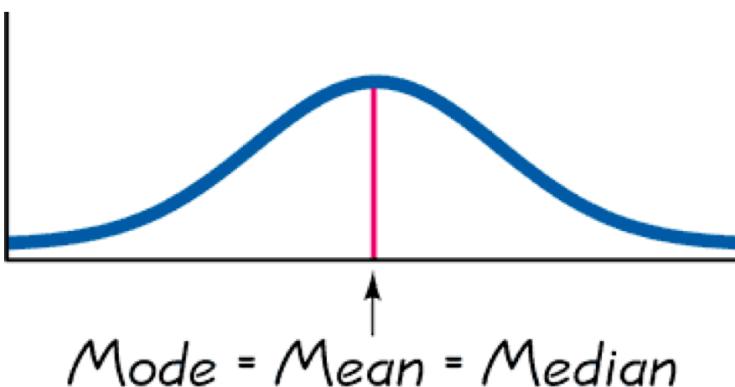
i	x
1	12
2	2
3	4
4	9
5	10
6	11
7	7
8	6
9	4
10	8

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = 7,3$$

\bar{x} = média

i = índice

n = número de elementos



Medidas de Dispersão

- Amplitude
- Desvio Médio
- Variância
- Desvio padrão
- Coeficiente de Variação
- Medidas centrais fornecem um resumo, mas não demonstram outra característica importante dos dados: a Dispersão. É preciso conhecer a variação dos dados para ter uma análise mais precisa.

Amplitude

- Corresponde à diferença entre o maior e o menor valor do conjunto de dados
- Diferença dos extremos
- Portanto muito sensível aos valores extremos
- $AMP = \max - \min$

Desvio Médio

- Média dos desvios
- 1) Calcula-se a média dos valores
- 2) Subtrai-se cada valor pela média
- 3) Calcula-se a média dos módulo da subtração
- $D = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$

Variância

- Média da soma dos quadrados dos desvios
- s^2 variância amostral
- σ^2 variância populacional

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

Variância

Exemplo: O diretor de fabricação deseja estimar a capacidade de uma linha de produção. Para tanto, ele levanta a quantidade de peças fabricadas durante 9 dias consecutivos e determina a sua média.

Dia	x	$(x_i - \bar{x})^2$
1	58	893,35
2	16	146,68
3	25	9,68
4	10	328,01
5	33	23,90
6	47	356,79
7	21	50,57
8	38	97,79
9	5	534,12

Porém, a produção em alguns dias está “distante” da média. No 9º dia, por exemplo, foram produzidas 5 peças.

Isso mostra que modelo da média não descreve plenamente a capacidade de fabricação da linha de produção.

Surge uma pergunta:

- Qual é a variação na produção observada?

Variância (S^2)= medida da dispersão (da variação) dos dados observados

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

$$\bar{x} = 28,11$$

$$S^2 = 271,21$$

Desvio Padrão

- Raiz quadrada da variância
- $s = \sqrt{s^2}$
- Quanto mais longe os valores da média maior é o desvio padrão
- Valores muito altos ou muito baixos influenciam no desvio padrão

Desvio Padrão

Dia	x	$(x_i - \bar{x})^2$
1	58	893,35
2	16	146,68
3	25	9,68
4	10	328,01
5	33	23,90
6	47	356,79
7	21	50,57
8	38	97,79
9	5	534,12

Exemplo: O diretor de fabricação deseja estimar a capacidade de uma linha de produção. Para tanto, ele levanta a quantidade de peças fabricadas durante 9 dias consecutivos e determina a sua média.

A variância é insuficiente porque é influenciada pelos valores que estão muito distantes da média.

O desvio padrão complementa a variância:

$$\sigma = \sqrt{S^2}$$

O desvio descreve o erro, no caso em qualquer valor fosse substituído pela média.

Assim:

$$\sigma = \sqrt{271,21} = 16,47$$

1. Uma conclusão já pode ser adotada:

Estimasse que o 10º dia produza $28,11 \pm 16,47$ unidades

Ou seja:

$$\underline{\min}_{10} = 11,64 \cong 12,00 \text{ unidades}$$

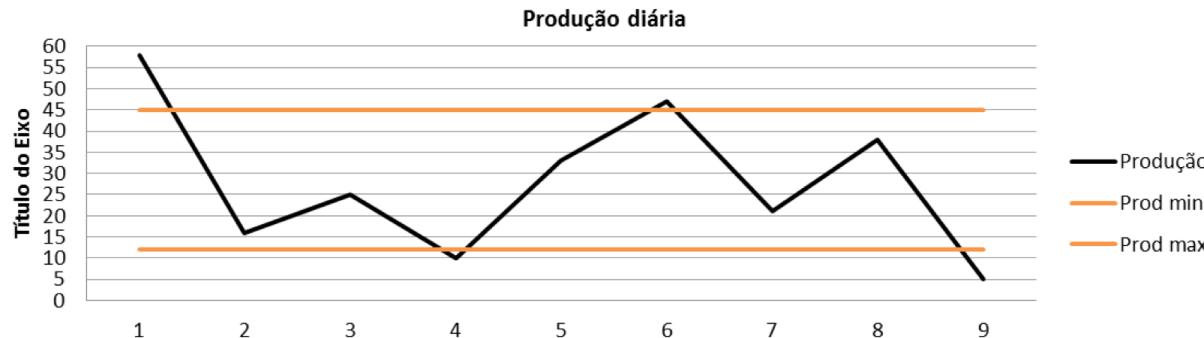
$$\overline{\max}_{10} = 44,58 \cong 45,00 \text{ unidades}$$

Desvio Padrão

2. Outra conclusão a ser adotada:

Se substituirmos qualquer um dos dias, o valor deve ser $28,11 \pm 16,47$ unidades. Determinando, assim, a AMPLITUDE 'mais provável'

3. To



mento:

Com essas informações, já é possível (entre outras coisas)

- a) Avaliar a estabilidade da produção (dos 9 dias, 4 estão fora de controle)
- b) Determinar metas para aumento ou redução das quantidades

Coeficiente de Variação

- Dispersão dos valores em relação à média.

$$CV = \frac{s}{\bar{x}} \cdot 100\%$$

- É muito útil para poder comparar conjuntos diferentes de dados

- Altura

- Média 1,64 m
- Desvio padrão 10 cm

$$CV = (10/164) \times 100 = 6,1\%$$

- Peso

- Média 60 kg
- Desvio Padrão 5 kg

$$CV = (5/60) \times 100 = 8,3\%$$

Medidas de Posição Relativa

- As medidas que vimos até agora podem não serem as mais adequadas para representarem os dados, Média e Desvio Padrão são muito sensíveis a valores extremos.
- Percentis
- Quartis
 - Boxplot

Percentis

- Dividem os dados em 100 grupos com cerca de 1% dos dados cada grupo;
- P_50 divide o conjunto de dados com 50% dos dados a baixo dele e 50% acima. É a mediana.
- Se a posição for intermediária, calcula-se a média

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

i = 1, ..., 99

n = número de elementos

Quartis

- Dividem os dados em 4 grupos com cerca de 25% dos dados cada grupo;
- Q_1 25% dos valores abaixo dele
- Q_2 divide o conjunto de dados com 50% dos dados a baixo dele e 50% acima. É a mediana.
- Q_3 75% dos valores abaixo dele
- Quartis são medidas resistentes.

$$Q_i = \frac{i(n+1)}{4},$$

i = 1, ..., 3
n = número de elementos

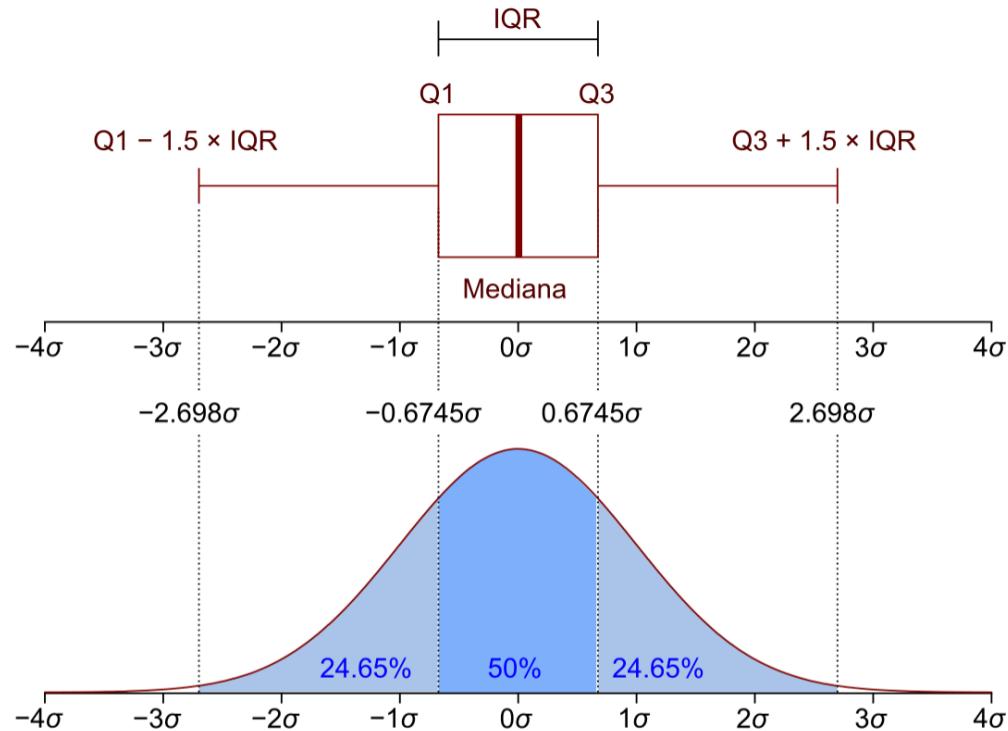
5 medidas

- Mínimo
 - 1 Quartil
 - 2 Quartil (Mediana)
 - 3 Quartil
 - Máximo
- 50% dos dados
- Com essas 5 medidas é possível ter uma boa perspectiva sobre a distribuição dos dados

Box Plot

- Uma representação bastante útil que tem por base as 5 medidas mais um limite superior (LS) e inferior (LI)
- $DQ = Q3 - Q1 = 50\% \text{ dos dados}$
- $LS = Q3 + 1,5 \times DQ$
- $LI = Q1 - 1,5 \times DQ$
- O 1,5 no cálculo garante que 99,3% dos dados esteja representado no gráfico segundo uma distribuição normal.

Box Plot



Quartis

aluno	notas
1	7
2	2
3	8
4	6
5	10
6	8
7	5
8	8
9	5
10	4
11	8
12	5
13	9

aluno	notas
2	2
10	4
7	5
9	5
12	5
4	6
1	7
3	8
6	8
8	8
11	8
13	9
5	10

Baixas

Mediana

Altas

aluno	notas
2	2
10	4
7	5
9	5
12	5
4	6
1	7
3	8
6	8
8	8
11	8
13	9
5	10

Mediana

Quartis

$$Q_{1/4} = 5$$

$$Q_{2/4} = 7$$

$$Q_{3/4} = 8$$

Quadrantes

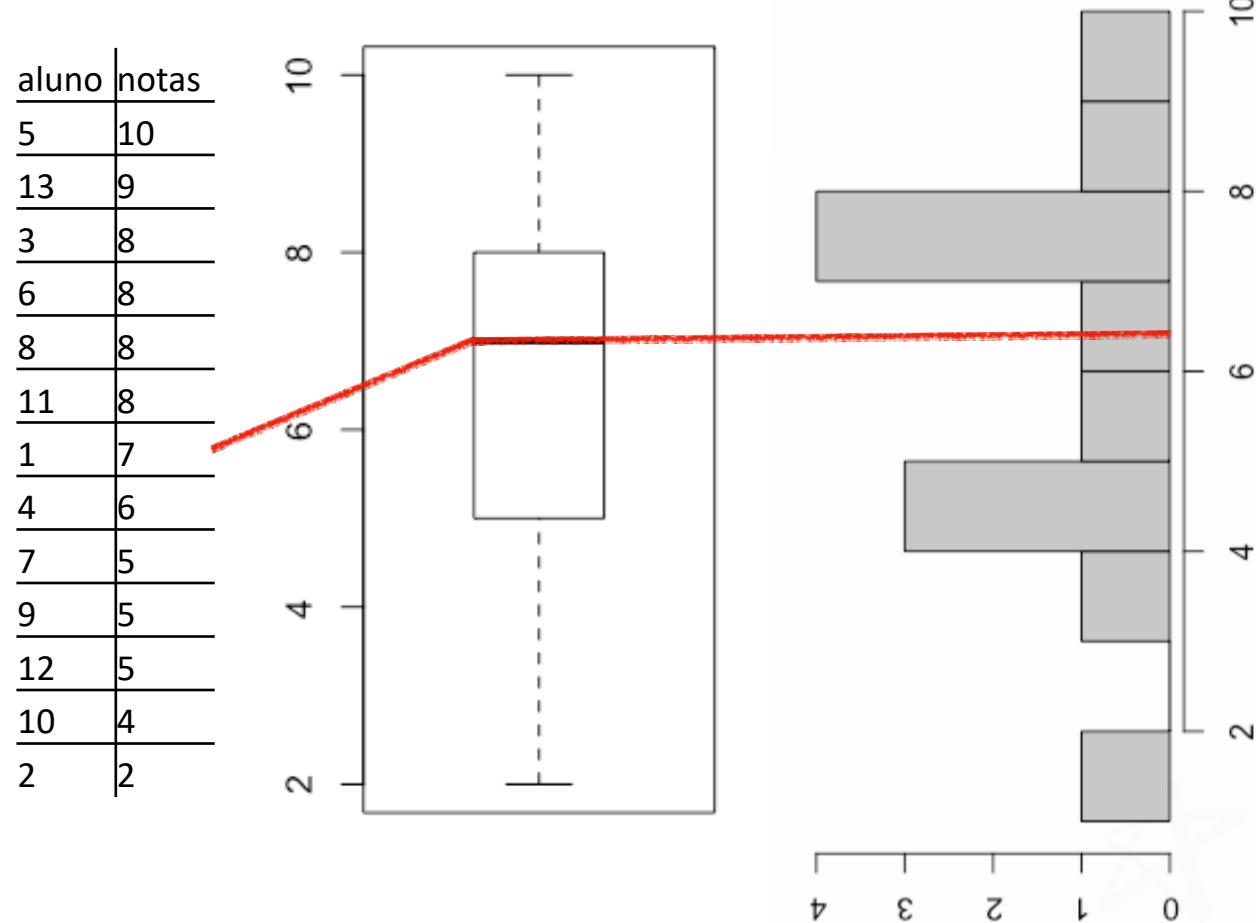
$$Q_1 = 0 < x \leq 5$$

$$Q_2 = 5 < x \leq 7$$

$$Q_3 = 7 < x \leq 8$$

$$Q_4 = 8 < x \leq 10$$

Boxplot



▲	dia	A	B	C	D
1	1	44	63	79	64
2	2	59	36	79	70
3	3	50	56	80	71
4	4	53	47	81	79
5	5	59	51	81	83
6	6	45	63	81	76
7	7	56	59	79	80
8	8	57	30	80	69
9	9	61	54	79	79
10	10	61	47	76	73

Média

A	54.5
B	50.6
C	79.5
D	74.4

Desvio P

A	6.293736
B	10.966616
C	1.509231
D	5.966574

Mediana

A	56.5
B	52.5
C	79.5
D	74.5

Quartis

▲	A	B	C	D
0%	44.00	30.00	76.00	64.00
25%	50.75	47.00	79.00	70.25
50%	56.50	52.50	79.50	74.50
75%	59.00	58.25	80.75	79.00
100%	61.00	63.00	81.00	83.00

