



Cruzeiro do Sul
Virtual
Educação a Distância

REPRESENTAÇÃO DE DISTRIBUIÇÕES

Prof. Ismar Frango



A Análise de Dados Exploratória – o que é?

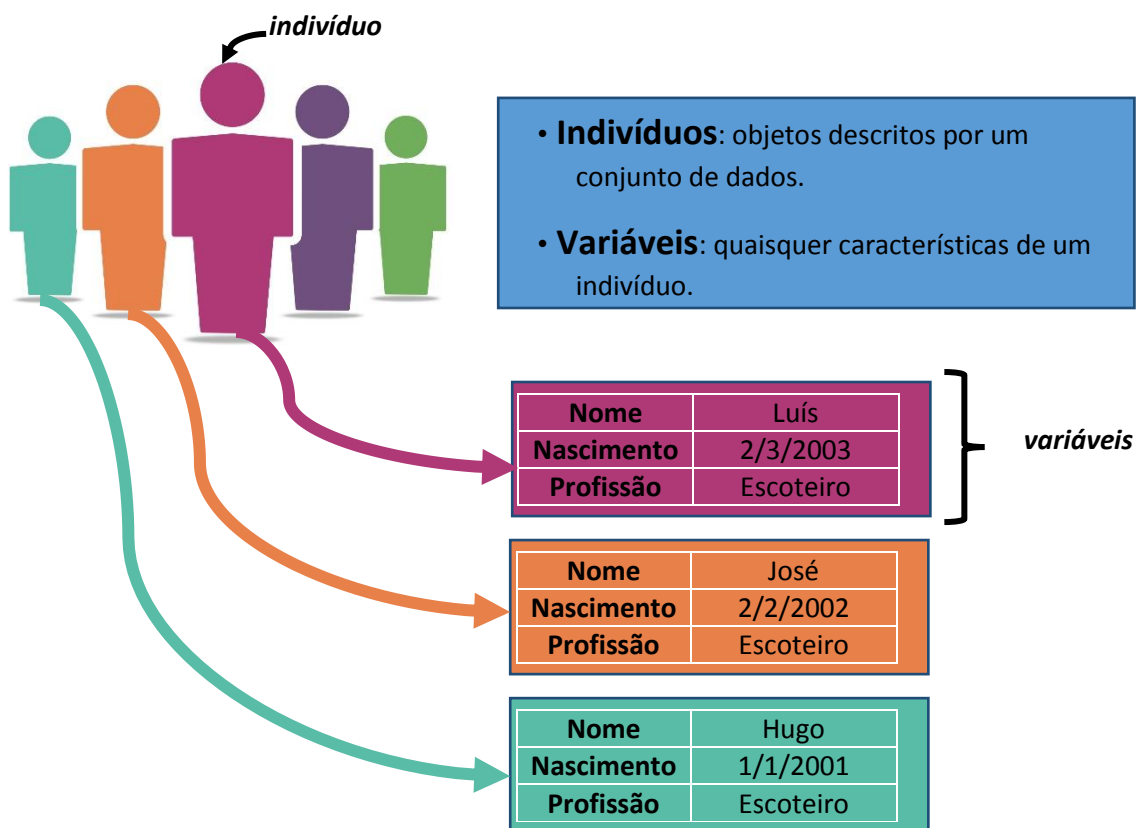
A Análise de Dados Exploratória consiste na aplicação de técnicas, ferramentas e conceitos da Estatística no tratamento de um conjunto de dados (aqui chamado de **dataset**), de forma a auxiliar na compreensão de suas principais características.



Saiba mais sobre *datasets* em:

https://pt.wikipedia.org/wiki/Conjunto_de_dados

Um tipo de *dataset* que é de interesse para essa área de conhecimento é aquele que contém **indivíduos** e **variáveis**.













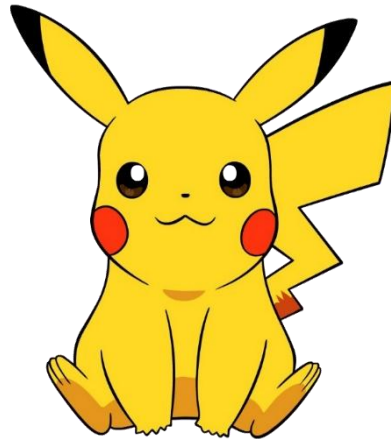
Fonte: Wikimedia Commons (Licença CC) + autor

Em geral, os *datasets* que contêm dados a respeito de indivíduos e variáveis são organizados assim: cada linha representa um indivíduo diferente, ao passo que cada coluna representa uma variável – nem sempre **tem** que ser assim, mas esta é a forma mais usual.

No exemplo a seguir, o *dataset* extraído do popular jogo Pokémon Go representa, em cada linha, um Pokémon. Cada coluna representa uma variável: nome do Pokémon, CP máximo, HP máximo, ataque, defesa e stamina (CP – *Combat Power*, HP – *Hit Points*, stamina, etc., são termos que caracterizam um Pokémon).

6 variáveis

#	POKÉMON	CP MÁXIMO ↓	HP MÁXIMO	ATAQUE	DEFESA	STAMINA
1	 Slaking	4.548	225	290	183	273
2	 Snorlax	3.355	261	190	190	320
3	 Blissey	3.219	410	129	229	510
4	 Ursaring	2.760	152	236	144	180
5	 Porygon2	2.546	144	198	183	170
6	 Tauros	2.488	128	198	197	150
7	 Kangaskhan	2.463	175	181	165	210
8	 Miltank	2.312	160	157	211	190
9	 Exploud	2.267	174	179	142	208
10	 Zangoose	2.214	125	222	124	146



10 indivíduos

Fonte: Pixabay (Licença CC) + Wikimedia Commons (Licença CC) + autor

Dado um conjunto de dados, devemos sempre fazer a análise **5W1H**. Ela consiste em realizar seis perguntas:



Saiba mais sobre 5W1H em:
<https://tinyurl.com/5w1h-explained>

Who? (Quem?)

- Quem ou quais são os indivíduos descritos pelo *dataset*? Quantos são esses indivíduos?

What? (O quê?)

- Que informações (variáveis) temos a respeito desses indivíduos? Quantas são essas variáveis? Em que unidades de medida são representadas?

Where? (Onde?)

- Em que contexto o *dataset* foi gerado? De onde provêm esses dados?

When? (Quando?)

- Quando os dados foram obtidos? O *dataset* ainda faz sentido?

Why? (Por quê?)

- Com que propósito o *dataset* foi gerado? Ele foi feito de maneira a nos permitir tirar conclusões a respeito apenas dos indivíduos nele representados ou ele contém indivíduos que são representativos de uma população maior?

How? (Como?)

- Como esses dados foram obtidos? É ético e legal utilizar esses dados?

Variáveis categóricas e quantitativas

Quando trabalhamos com Análise de Dados Exploratória, é essencial sabermos a diferença entre variáveis categóricas e quantitativas.

Variáveis categóricas:

servem para categorizar o indivíduo, situando-o em um determinado grupo

Variáveis quantitativas:

têm valores numéricos, de acordo com alguma unidade de medida



Modelo	Combustível	Cilindros	Km/l cidade	Km/l estrada
Toyota Prius 1.8	Híbrido	4	18,9	16
VW up! TSI 1.0	Gasolina	3	14,3	16,3
Fiat Mobi 1.0	Gasolina	4	13,7	16,1
Chevrolet Prisma 1.0	Gasolina	4	13,1	15,8
Land Rover Evoque 2.0 SE TD4	Diesel	4	11,9	15,8



Nem toda variável numérica é quantitativa.
Neste exemplo, a quantidade de cilindros

Fonte: Pixabay (Licença CC) + autor

Dataset: Internet + Revista Auto Esporte (<https://revistaautoesporte.globo.com/Noticias/noticia/2018/05/os-20-carros-mais-economicos-do-brasil.html>)

Variáveis categóricas

As variáveis categóricas (que também são chamadas de **qualitativas**) representam uma classificação dos indivíduos em categorias. Elas podem ser **nominais** ou **ordinais**, quando há ou não uma relação de ordem.

Exemplos:

- Gênero
- Método de pagamento
- Fumante / Não fumante
- Cor dos olhos

Exemplos:

- Estágio de uma doença (inicial → intermediário → terminal)
- Escolaridade (fundamental → médio → superior)
- Temperatura de um chuveiro (frio → morno → quente)

Representações gráficas bastante comuns para variáveis categóricas são: os gráficos de colunas (ou barras) e os gráficos de setores (popularmente conhecidos como “de pizza” ou, em inglês, *pie charts*).

Vamos ver um exemplo a seguir:

Clube	% de torcedores	Projeção / Brasil
Flamengo	18%	36.475.167
Corinthians	14%	28.369.574
São Paulo	8%	16.211.185
Palmeiras	6%	12.158.389
Vasco	5%	10.131.991
Grêmio	4%	8.105.592
Cruzeiro	3%	6.079.194
Santos	3%	6.079.194
Internacional	3%	6.079.194
Atlético Mineiro	2%	4.052.796
Botafogo	2%	4.052.796
Fluminense	2%	4.052.796
Bahia	1%	2.026.398
Vitória	1%	2.026.398
Outros	5%	10.138.428
Nenhum ou Fora da faixa	23%	46.729.470
Total	100%	202.768.562

Fonte: Wikimedia Commons (Licença CC) + autor
Dataset: Datafolha – pesquisa realizada entre 3 e 5 de junho de 2014

Em pesquisa realizada pelo instituto Datafolha no dia 3 e 5 de junho de 2014, foram entrevistadas 4.337 pessoas em 207 municípios brasileiros. A faixa etária escolhida foi acima de 16 anos. Perguntou-se o time predileto de cada um dos entrevistados. Os resultados finais da pesquisa estão na tabela acima. Pesquisas como essas, considerando uma **amostra** podem ser extrapolados para toda a **população**, se realizadas seguindo um conjunto de critérios.

Note que o time escolhido representa uma **categoria** – quando os entrevistados responderam à pesquisa, eles foram **classificados** de acordo

com o time predileto. Provavelmente, o instrumento de pesquisa do Datafolha pode ter sido algo parecido com isto:

Nome	Gênero	Idade	Profissão	Renda	Atlético	Bahia	Corinthians	Cruzeiro	Flamengo
José das Couves	M	55	Advogado	20000		X			
Maria das Graças	F	34	Engenheira	15000			X		
João dos Milages	M	19	Programador	10000					X

Ou isto:

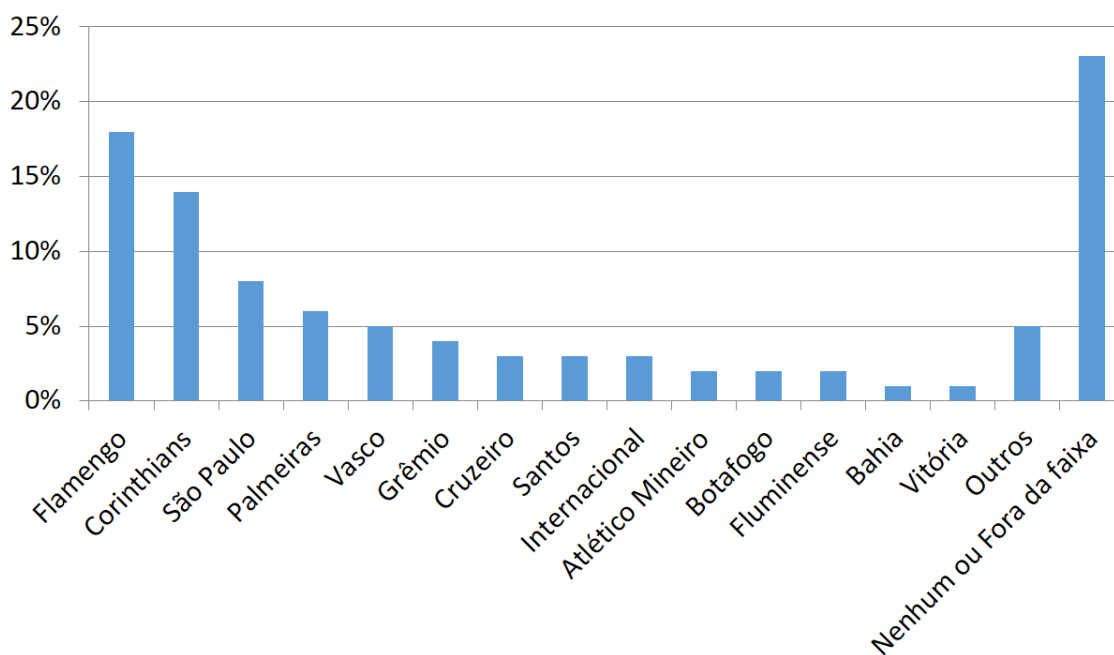
Fonte: autor

Nome	Gênero	Idade	Profissão	Renda	Time
José das Couves	M	55	Advogado	20000	Bahia
Maria das Graças	F	34	Engenheira	15000	Corinthians
João dos Milages	M	19	Programador	10000	Flamengo

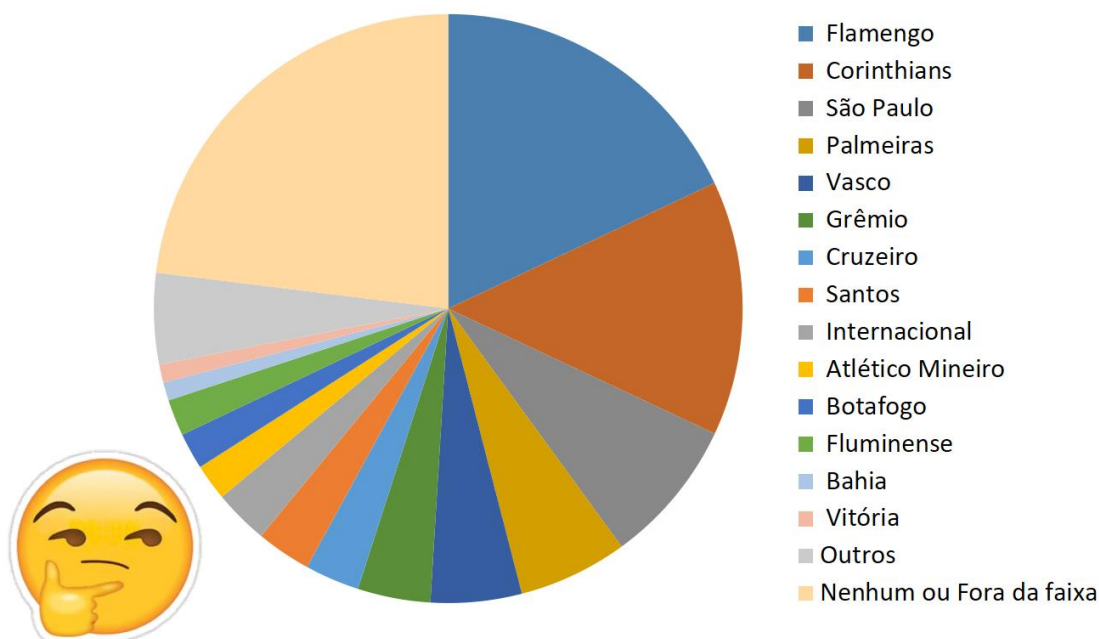
Enquanto o primeiro instrumento tem a limitação de ter que representar todos os times existentes, um por coluna (e algumas colunas possivelmente serem mais preenchidas que outras), o segundo instrumento pode ser passível de erro de preenchimento manual, em especial nas variáveis “Profissão” e “Time”.

Veja que quase todas as variáveis, em ambos os casos, são categóricas (exceto idade e renda). Uma forma de **representar quantitativamente** os resultados relacionados a uma variável categórica é por meio de gráficos como o de colunas (ou barras):

% Torcedores por clube



Uma outra representação possível seria com o gráfico de setores. Porém, nem sempre este gráfico se demonstra adequado para todo tipo de distribuição.

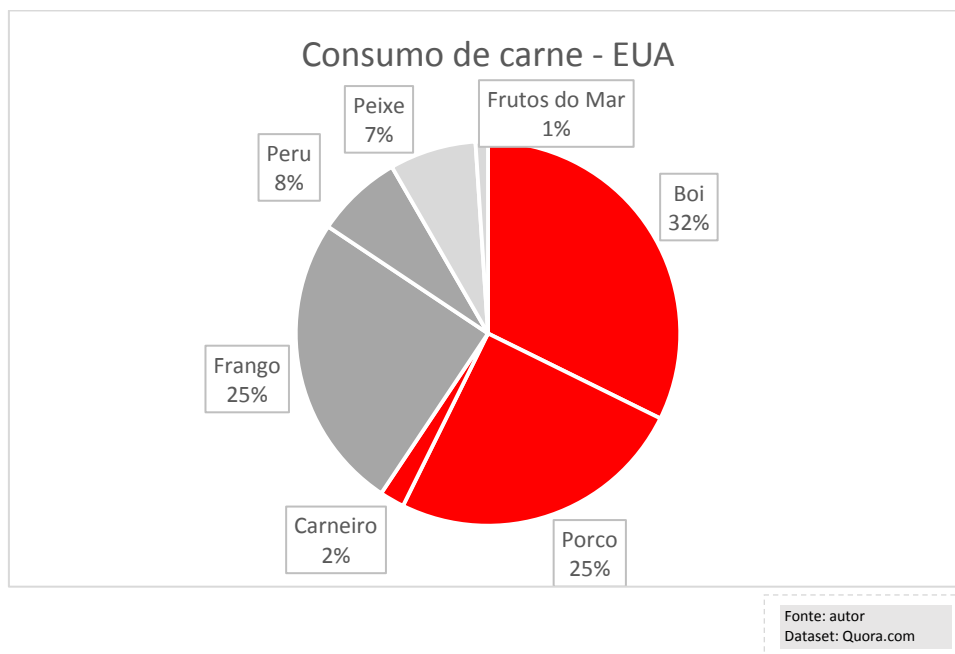


Veja que este gráfico de setores tem uma série de problemas:

- Algumas cores utilizadas são muito próximas. Como diferenciar, por exemplo, os torcedores do Corinthians e os do Santos?
- Nós, humanos, não conseguimos diferenciar muito bem entre ângulos que têm valores próximos. Quem tem mais torcedores, o Cruzeiro ou o Santos?
- O fato dos valores não estarem exibidos para cada categoria faz com que a informação esteja incompleta.
- Gráficos de setores são bons para comparar **uma** parte com o todo



- Gráficos de setores são bons para comparar **grupos de partes** com o todo (no exemplo, é facilmente visível que a maior parte de carne consumida nos EUA é carne vermelha)



O problema é que na maior parte do tempo, quando lidamos com variáveis categóricas, o que nos interessa não são **partes comparadas ao todo**, mas sim as **partes comparadas às outras partes**. Nesses casos, **gráficos de barras** (ou colunas) **são melhores**, pois são mais flexíveis que os gráficos de setores, já que com eles, é possível comparar qualquer conjunto de variáveis categóricas que sejam medidas na mesma unidade.

Variáveis quantitativas

Variáveis quantitativas são aquelas que geralmente podem ter muitos valores, inclusive infinitos, não permitindo a construção de categorias para classificar os indivíduos. Nesse sentido, o gráfico bastante utilizado na representação da distribuição de uma variável quantitativa é o **histograma** – para ele ser feito, é necessário agrupar valores próximos em intervalos (também chamados de classes ou *bins*). O histograma representa **frequências** – isto é, a quantidade de ocorrências de valores pertencentes a um intervalo.

Tomemos, como exemplo, uma distribuição com diversos indivíduos e suas idades (em anos):

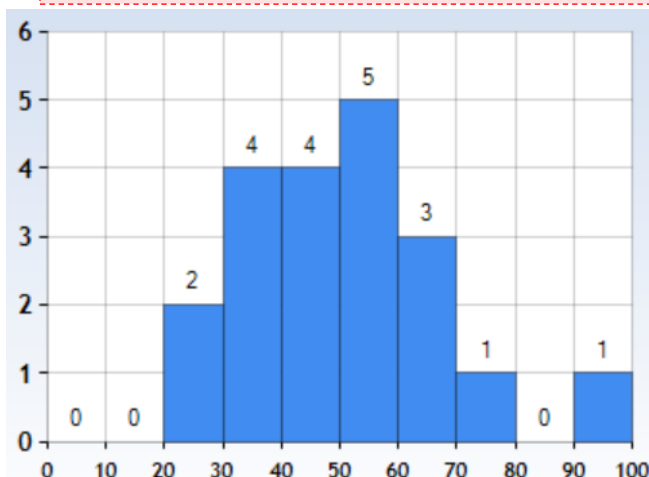
36	25	38	46	55	68	72	55	36	38
67	40	22	48	91	46	52	61	58	55

Para representarmos o histograma desta distribuição, temos que definir o tamanho do intervalo. Vamos trabalhar com um intervalo de tamanho 10, e distribuir as ocorrências em seus intervalos respectivos. Em seguida, obtém-se o **histograma**.

Intervalo	Valores no intervalo	Frequência
0-10	-	0
10-20	-	0
20-30	25,22	2
30-40	36,38,36,38	4
40-50	46,40,48,46	4
50-60	55,55,52,58,55	5
60-70	68,67,61	3
70-80	72	1
80-90	-	0
90-100	91	1

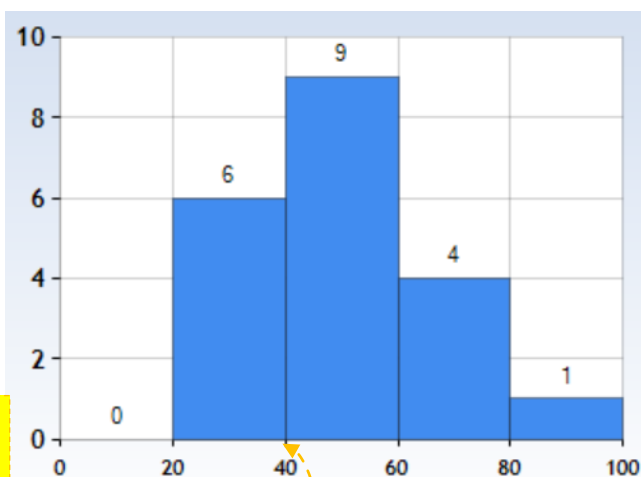


Uma ferramenta online para gerar histogramas está em:
<https://www.socscistatistics.com/descriptive/histograms/>



Fonte: autor

Perceba que, quando se altera o tamanho do intervalo, o formato do histograma também é alterado. Por exemplo, com o mesmo *dataset*, vamos representar um histograma com intervalo de 20, ao invés de 10.



Fonte: autor



Convencionou-se que um valor na fronteira entre intervalos vai sempre para o intervalo da direita. Por exemplo, a idade de 40 anos ficou dentro do intervalo 40-60, não no 20-40.

Fonte: PXHere (CC)



Como construir um histograma?

1. Escolha o tamanho do intervalo.
2. Divida os dados em seus respectivos intervalos.
3. Conte os indivíduos em cada classe.
4. Desenhe o histograma – à mão, ou usando a ferramenta computacional de sua preferência.

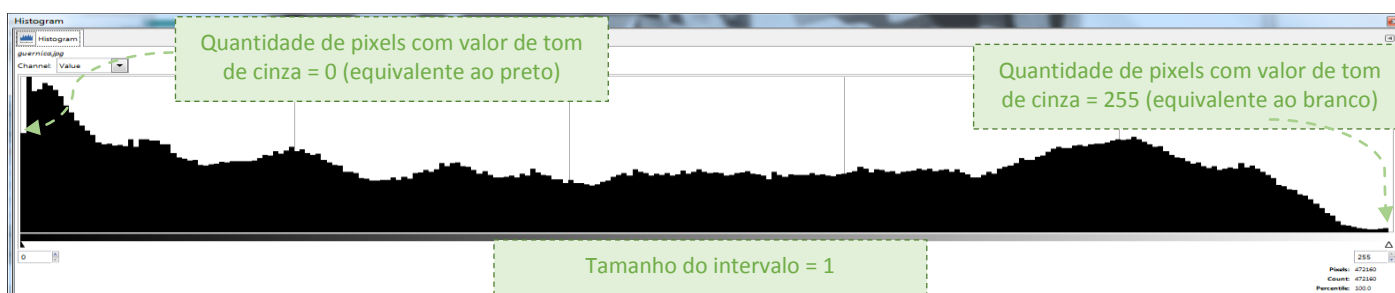
O endereço a seguir é uma boa ferramenta online para confecção de histogramas:

<https://www.socscistatistics.com/descriptive/histograms/>

Há diversas aplicações para histogramas em muitas áreas do conhecimento. Por exemplo, o famoso quadro Guernica, do pintor espanhol Pablo Picasso, retratando os horrores da Guerra Civil Espanhola, foi pintado todo em tons de cinza. Um histograma gerado por uma ferramenta de edição de imagens analisou a frequência (quantidade) de pixels com cada um dos 256 diferentes tons de cinza a partir de uma fotografia do quadro:



Pablo Picasso. Guernica, 1937.
Madri, Museu Rainha Sofia



Fonte: autor

Embora **histogramas** se pareçam com gráficos de barras, há diferenças importantes:

- Um histograma apresenta a distribuição de uma variável quantitativa; gráficos de barras, em geral, são usados para variáveis categóricas;
- Em geral, as barras do histograma são apresentadas lado a lado, sem espaçamento entre elas;
- **O eixo horizontal de um histograma** deve referenciar as unidades de medida da variável.
- **O eixo vertical de um histograma** identifica a frequência (quantidade de ocorrências em um intervalo), sem escala de medida.

A Mediana

Uma forma de descrever numericamente o centro de uma distribuição é por meio de sua mediana. A mediana é o valor que divide um conjunto de dados **ordenado** em duas partes iguais: **metade** das observações está valores **abaixo** da mediana e **metade, acima** dela.

Considere uma distribuição de n indivíduos; a mediana está na posição :

$$\frac{n + 1}{2}$$

Por exemplo, na seguinte distribuição:

1, 3, 3, **6**, 7, 8, 9

Mediana

$$n = 7 \rightarrow (n+1)/2 = 4^{\text{a}} \text{ posição}$$

E nesta distribuição:

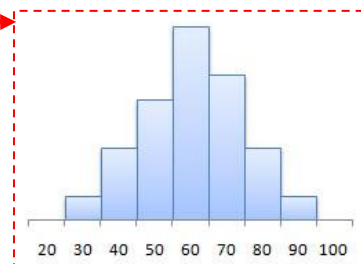
1, 2, 3, **4**, **5**, 6, 8, 9

Mediana

$$n = 8 \rightarrow (n+1)/2 = "4,5"^{\text{a}} \text{ posição}$$

Quando temos um n par, calcula-se a mediana como sendo a média dos dois valores que estão à esquerda e à direita do ponto central (no último exemplo, o valor da mediana seria $(4+5)/2 = 4,5$).

Uma distribuição é **simétrica** se os lados direito e esquerdo do histograma, a partir da mediana, são **aproximadamente** iguais.



Uma distribuição é **assimétrica à direita** se o lado direito do histograma (a metade das observações com valores maiores), contando a partir da mediana, se prolonga mais que o lado esquerdo. Uma distribuição é **assimétrica à esquerda** quando o oposto acontece.

O Diagrama de Ramo e folhas (*Stemplot*)

O diagrama de ramo e folhas (ou *stemplot*, em inglês), ainda que menos utilizado do que o histograma, é uma outra forma de visualizar uma distribuição. Por exemplo, utilizando a mesma distribuição de idade de um exemplo anterior:

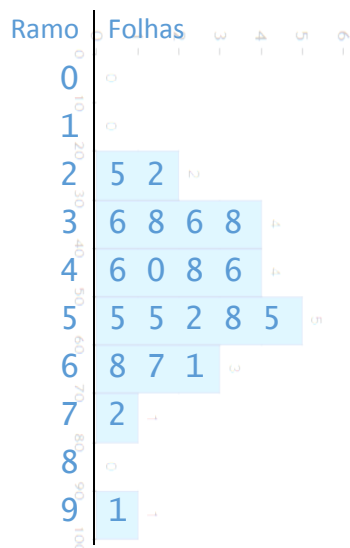
36	25	38	46	55	68	72	55	36	38
67	40	22	48	91	46	52	61	58	55

Uma visualização desta distribuição como um diagrama de ramo-e-folhas é feita de maneira parecida com o histograma: escolhe-se um “ramo” (no caso desta distribuição, escolheremos as dezenas) e em seguida distribuem-se as “folhas”:

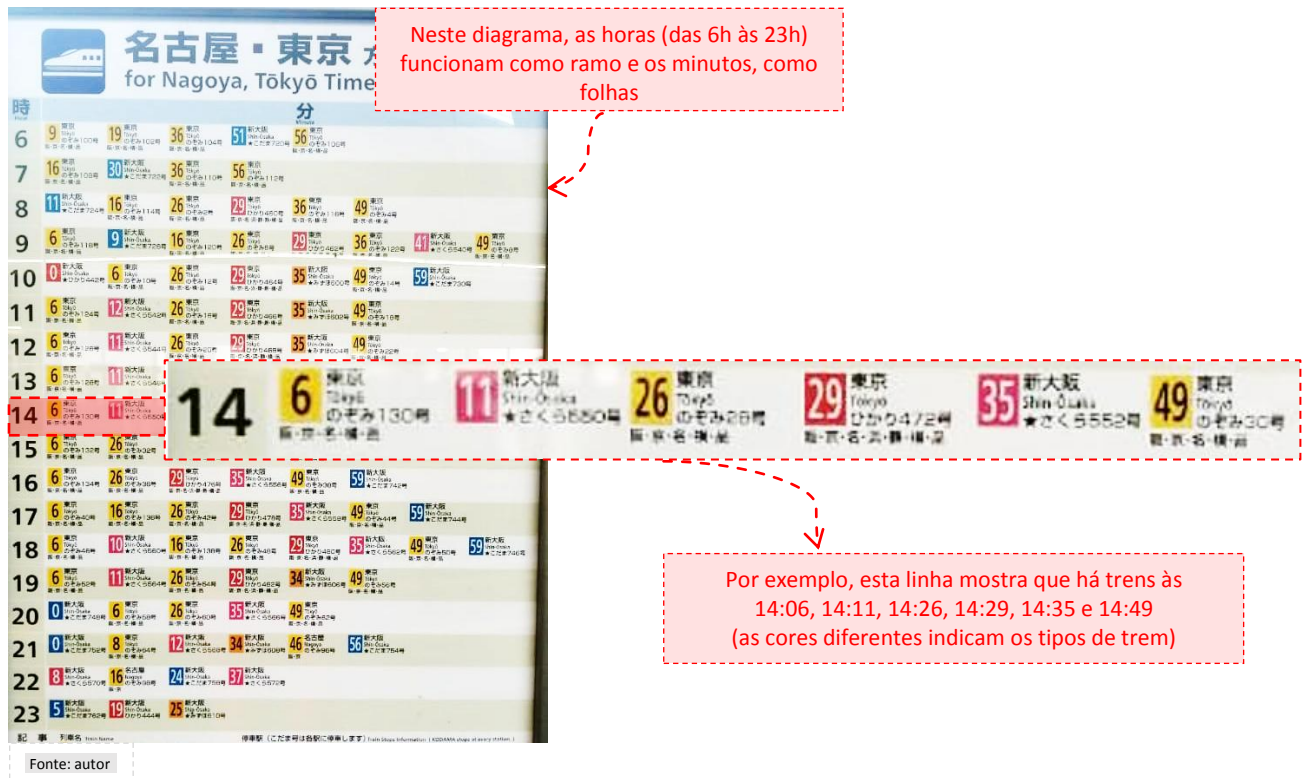
Ramo	Folhas
0	
1	
2	5 2
3	6 8 6 8
4	6 0 8 6
5	5 5 2 8 5
6	8 7 1
7	2
8	
9	1

O número 58 é representado na linha do **ramo 5** (equivalente a 5 dezenas) e a **folha de valor 8** (equivalente a 8 unidades)

Note que este diagrama, visualmente, remete ao histograma desta mesma distribuição (com intervalo 10), rotacionado:

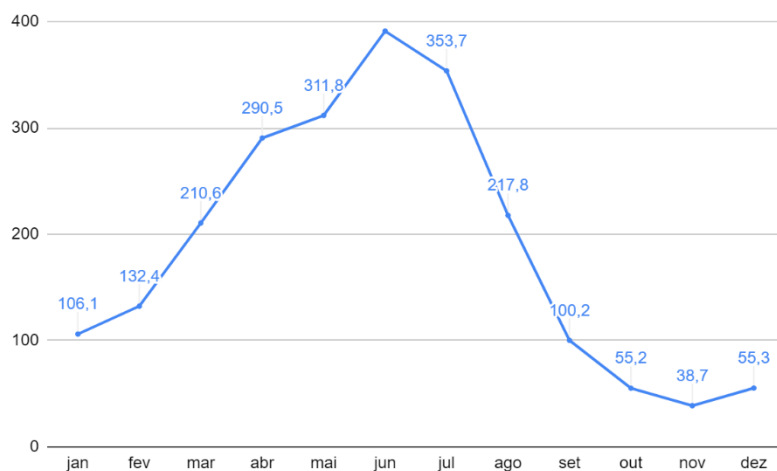


Algumas culturas utilizam esta representação de maneira mais natural que a nossa. Veja o exemplo de uma foto tirada em uma estação de trem em no Japão, que um diagrama de ramo-e-folhas é usado para indicar os horários dos trens:



Gráficos temporais

Um **gráfico temporal** representa o valor de cada observação em relação ao tempo em que foi obtida, como no exemplo a seguir, que mostra a precipitação (mm de chuva) em Recife, considerando a média mensal:



Fonte: autor
Dataset: Instituto Nacional de Meteorologia (INMET), dados de 1981-2010

Em um gráfico temporal, geralmente o tempo fica no eixo das abscissas (x) e a variável observada, no eixo das ordenadas. É comum, como no exemplo, conectar as observações com uma linha.

Um gráfico temporal pode apresentar **ciclos**, ou seja, um padrão geral, bem como **tendências**.



Representações numéricas de distribuições

Após estudarmos algumas representações gráficas de distribuições, vamos nos debruçar sobre como representar distribuições por meio de números.



O conteúdo a seguir tem um caráter revisional, já que o mesmo geralmente é coberto por uma disciplina de Probabilidade e Estatística básica. Porém, como ele é importante para a Análise Estatística de Dados, ele é retomado aqui.

Média

A medida mais comum de **centro de uma distribuição** é a **média**.

Dados os valores de n observações x_1, x_2, \dots, x_n de uma distribuição, a média (representada por μ ou \bar{x}) dessas observações é dada por:

(Há várias formas de cálculo de média: ponderada, harmônica, etc. Aqui estamos trabalhando com a mais comum delas por enquanto, a **média aritmética** – vamos chamá-la simplesmente de média nesse contexto).

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^n x_i$$

Dizemos que a média não é uma medida de centro resistente já que ela pode ser afetada pela presença de valores não-característicos em uma distribuição, também conhecidos como *outliers*.

Por exemplo, qual a média da distribuição de idades que mostramos em exemplos anteriores?

Média = 50,45

E se tirarmos a observação de valor 91?

Média = 45,9

Note o quanto uma única observação (no caso, o valor 91) interfere na média!

Ramo	Folhas
0	
1	
2	5 2
3	6 8 6 8
4	6 0 8 6
5	5 5 2 8 5
6	8 7 1
7	2
8	
9	1

Moda

A moda de uma distribuição é o valor (ou os valores) mais comum, ou seja, o que mais aparece (m) em uma distribuição, que pode ter **uma, duas ou mais modas** – sendo classificada como **monomodal**, **bimodal** ou **multimodal**.



Assista ao *gameplay* de um jogo educativo sobre Moda:
<https://youtu.be/lpqeC5-8npQ>

Ramo	Folhas
0	
1	
2	5 2
3	6 8 6 8
4	6 0 8 6
5	5 5 2 8 5
6	8 7 1
7	2
8	
9	1

Por exemplo, na nossa distribuição de idades:

A idade que ocorre com mais frequência é 55 anos (aparece 3 vezes), logo:

moda = 55

Amplitude

A amplitude A de uma distribuição é a diferença entre o maior e o menor valor (Max e Min) de uma distribuição:

$$A = \text{Max} - \text{Min}$$

Por exemplo, na nossa distribuição de idades:

Ramo	Folhas
0	
1	
2	5 2
3	6 8 6 8
4	6 0 8 6
5	5 5 2 8 5
6	8 7 1
7	2
8	
9	1

Max = 91

Min = 22

$A = 91 - 22 = 69$

O que acontece se eliminarmos a observação com valor 91?

Max = 72

$A = 72 - 22 = 50$

Assim como a média, a amplitude é muito sensível à existência de *outliers*.

Variância

A **variância s^2** de um conjunto de observações é uma medida bastante importante, calculada como a seguir, para uma distribuição com n valores:

1. Calcule o quadrado da **diferença** de cada valor observado para a média
2. Some este valor
3. Divida por $n-1$

Também se usa o termo "desvio"

Ou, resumidamente:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

No nosso exemplo:

\bar{x}	50,45
$\sum_{i=1}^n (x_i - \bar{x})^2$	5262,95
s^2	276,99

média

soma dos quadrados dos desvios

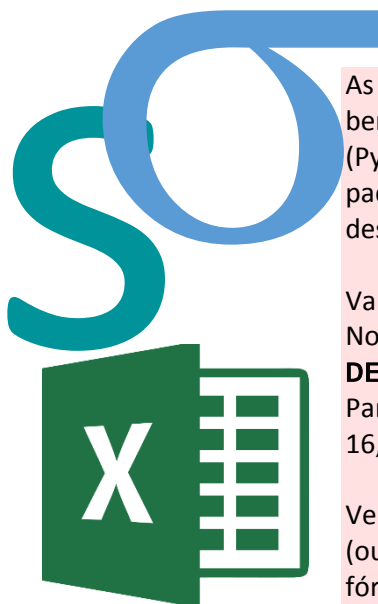
variância

Desvio-padrão

Bem mais utilizado que a **variância s^2** , o desvio-padrão mostra o quão dispersos (espalhados) os dados se encontram em relação à média. Logo, o desvio-padrão é uma medida de dispersão. Ele é calculado basicamente como a raiz quadrada de s^2 .

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Usamos **s** é usado para representar o **desvio padrão** de uma amostra. O símbolo σ (**sigma**) é frequentemente usado para representar o **desvio padrão** de uma população, com um cálculo ligeiramente diferente (no lugar de $n-1$, usamos n).



As planilhas de cálculo (Microsoft Excel, Google Spreadsheet, etc), bem como as linguagens de programação para manipulação de dados (Python, R, etc) possuem funções prontas para o cálculo do desvio-padrão. Porém, é importante saber que há duas formas de cálculo do desvio-padrão, uma para a **amostra** e outra para a **população**!

Vamos tomar o Microsoft Excel em português como exemplo: No Excel, há duas funções para cálculo de desvio-padrão: **DESVPAD.A** e **DESVPAD.P** (em inglês: **STDEV.S** e **STDEV.P**). Para nosso exemplo acima, o valor de **DESVPAD.A** (amostra) é 16,64, enquanto **DESVPAD.P** é 16,22.

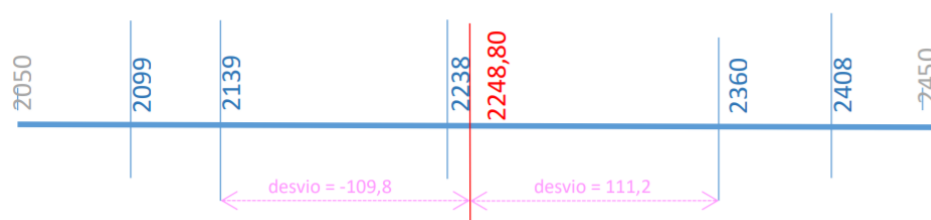
Versões mais antigas da planilha possuem apenas a função **DESVPAD** (ou **STDEV**), que é equivalente à **DESVPAD.A** – que é, em geral, a fórmula usada mais frequentemente.

Vamos trabalhar com um exemplo completo:

Uma busca por uma geladeira em um site comparador de preços retornou os seguintes cinco melhores resultados no e-commerce (em R\$):

Loja 1	Loja 2	Loja 3	Loja 4	Loja 5
2238	2360	2139	2408	2099

$$\text{Cálculo da média: } \bar{x} = \frac{2238+2360+2139+2408+2099}{5} = 2248,80$$



Fonte: Pixabay (Licença CC)

Lojas i	Observações x_i	Desvios $x_i - \bar{x}$	Desvios ² $(x_i - \bar{x})^2$
1	2238	-10,8	116,64
2	2360	111,2	12365,44
3	2139	-109,8	12056,04
4	2408	159,2	25344,64
5	2099	-149,8	22440,04

A **variância** é a soma dos quadrados dos desvios, dividida pelo número de observações, que são cinco, menos 1:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2}{4} = 18080,7$$

E o **desvio-padrão** é a raiz quadrada da variância:

$$s = \sqrt{18080,7} \cong 134,46$$

Medidas de variabilidade: Quartis

Quartis são três valores que dividem uma distribuição com n valores em quatro partes.

- O 1º quartil Q1 é a mediana das observações localizadas à esquerda da localização da mediana:

$$\text{Posição do Q1} = \frac{n+1}{4}$$

- O 2º quartil Q2 é a própria mediana;
- O 3º quartil Q3 é a mediana das observações localizadas à direita da localização da mediana:

$$\text{Posição do Q} = 3 \frac{n+1}{4}$$

- A amplitude inter-quartis **AIQ** é a diferença entre o 3º. e o 1º. quartis:

$$\text{AIQ} = \text{Q3} - \text{Q1}$$

Uma boa ideia da amostra é dada pelos cinco valores seguintes (escritos em ordem crescente), chamados de “Resumo dos cinco números”:

Min Q1 M Q3 Max

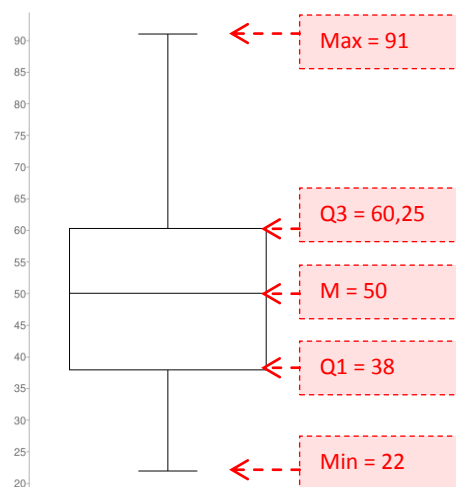
Uma representação gráfica adequada para visualizar uma distribuição com destaque para esses valores é o *boxplot*, ou gráfico de caixa. Vejamos um exemplo com o *dataset* de idades já utilizado em outros exemplos, porém agora ordenado de maneira crescente:

1	22
2	25
3	36
4	36
5	38
6	38
7	40
8	46
9	46
10	48
11	52
12	55
13	55
14	55
15	58
16	61
17	67
18	68
19	72
20	91

Posição do Q1: $(20+1)/4 = 5,25$
(entre 5ª e 6ª posições)

Posição da mediana: $(20+1)/2 = 10,5$
(entre 10ª e 11ª posições)

Posição do Q3: $3 \cdot (20+1)/4 = 15,75$
(entre 15ª e 16ª posições)



O cálculo do valor de Q3, neste exemplo, é feito da seguinte forma: a diferença entre os valores das posições 15 e 16 é $61-58=3$; como a posição de Q3 é “15,75ª”. Calcula-se então $0,75 \cdot 3=2,25$ e soma-se AP valor de 58, que dá 60,25. No caso de Q1, como os valores entre as posições 5 e 6 são iguais, o valor de Q1 é exatamente o mesmo.



Você pode gerar seu *boxplot* online em ferramentas como:
<http://www.alcula.com/calculators/statistics/box-plot/>

Um ponto importante a se ressaltar em planilhas eletrônicas é a existência de dois métodos diferentes para cálculo dos quartis. No Microsoft Excel em português, por exemplo, a função QUARTIL.EXC localiza os quartis exatamente como descrito neste material; já a função QUARTIL.INC utiliza $n-1$ ao invés de $n+1$ no cálculo das posições levando a valores diferentes.

Regra 1,5 AIQ para detecção de *outliers*

Há uma regra para definir se um valor é um *outlier* em uma distribuição, que é a Regra 1,5AIQ: se o valor está a uma distância de mais de $1,5 \times \text{AIQ}$ abaixo do primeiro quartil ou acima do terceiro quartil, ele pode ser considerado um *outlier*. Em um boxplot, é comum representar os *outliers* com asteriscos (*) e considerar como valores máximo e mínimo de uma distribuição apenas valores que não sejam *outliers*.

Veja que, em nosso exemplo, a maior idade (91) na verdade não é um *outlier*, de acordo com a Regra 1,5AIQ: a AIQ de nosso exemplo é:

$$Q3 - Q1 = 60,25 - 38 = 22,25$$

$$1,5\text{AIQ} = 1,5 * 22,25 = 33,375$$

Ou seja, para ser um *outlier*, o valor da observação deve ser maior do que:

$$Q3 + 1,5\text{AIQ} = 60,25 + 33,375 = 93,625$$

ou menor do que

$$Q1 - 1,5\text{AIQ} = 38 - 33,375 = 4,625$$



Para saber mais, leia os capítulos 1 e 2 do e-book:

MOORE, David S.; NOTZ, William I.; FLINGER, Michael A. A Estatística Básica e sua Prática. 7 ed. Rio de Janeiro: LTC, 2017 – Capítulo 02