SENAI

Ciência de Dados (CIDA)

Aula 2 –Estatística Descritiva

Prof.: Hugo S. Idagawa

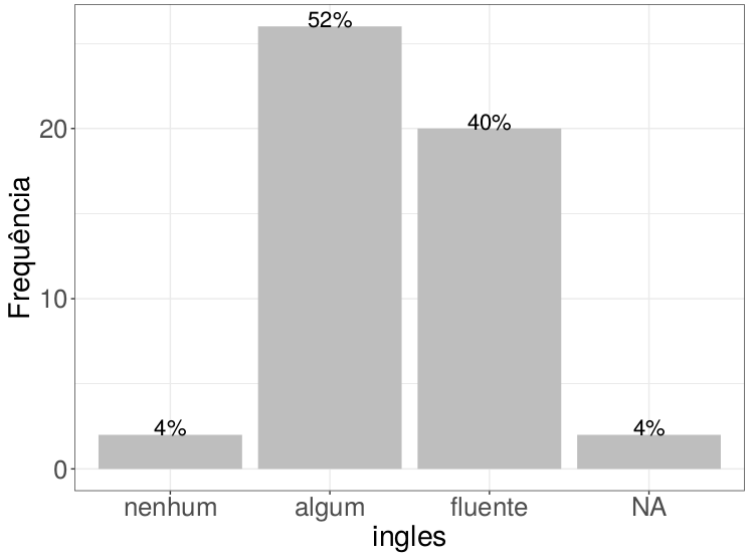
Distribuição de Frequência

- Após a coleta de dados, temos em mãos uma tabela de dados que contém informações sobre uma variável da amostra. Uma etapa importante da análise descritiva é a criação de gráficos que permitem visualizar e compreender melhor o fenônemo que está sendo estudado.

- Além do diagrama de caixas (Boxplot), os gráficos de frequência permitem visualizar a distribuição das variáveis e servem tanto para variáveis qualitativas (não-numéricas) quanto para variáveis quantitativas (numéricas).

Fluência em inglês	Frequência observada	Frequência relativa (%)	Frequência acumulada (%)
nenhuma	2	4	4
alguma	26	54	58
fluente	20	42	100
Total	48	100	

Obs: dois participantes não forneceram informação.



(Exemplo de distribuição para uma variável qualitativa)

Distribuição de Frequência

- No caso de variáveis qualitativas (especialmente no caso de variáveis contínuas), ao construir um gráfico de frequência, podemos acabar com uma variação de frequências muito pequenas deixando de cumprir o objetivo de resumir os dados. Assim, para contornar essa situação, é comum agrupar os valores em classes e em seguida obter a frequência de cada classe.

Exemplo: tabela de um questionário sobre salário respondido por 50 alunos (apenas os 30 primeiros estão apresentados).

Aluno	1	2	3	4	5	6	7	8	9	10
Salário (R$)	3500	1800	4000	4000	2500	2000	4100	4250	2000	2400

Aluno	11	12	13	14	15	16	17	18	19	20
Salário (R$)	7000	2500	2800	1800	3700	1600	1000	2000	1900	2600

Aluno	21	22	23	24	25	26	27	28	29	30
Salário (R$)	3200	1800	3500	1600	1700	2000	3200	2500	7000	800

Distribuição de Frequência

- Ao observar a tabela anterior, é possível verificar que alguns valores de salários apenas aparecem uma única vez. Assim, para melhorar a visualização, podemos agrupar os salários em classes de R$1500,00 de largura.

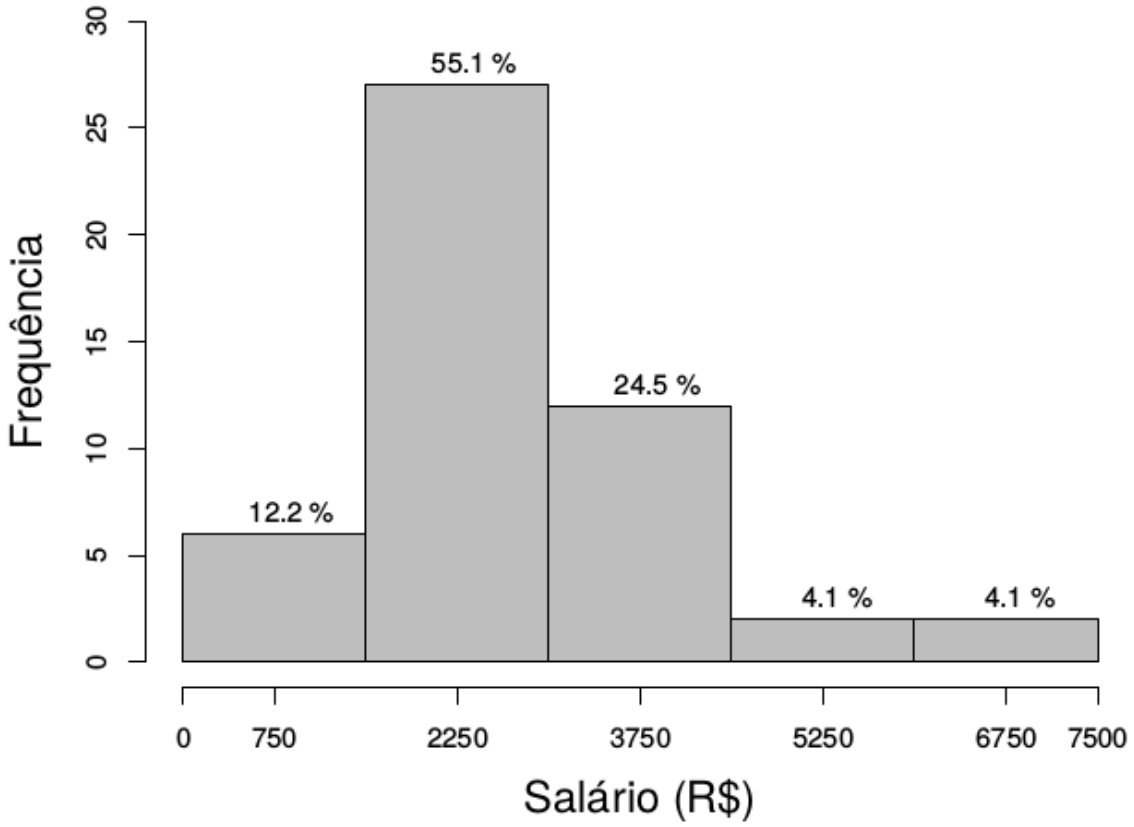
- Em seguida, podemos reconstruir a tabela anterior da seguinte maneira:

Classe de salário (R$)	Frequência observada	Frequência relativa (%)	Frequência relativa acumulada (%)
$0 \vdash 1500$	6	12,2	12,2
$1500 \vdash 3000$	27	55,1	67,3
$3000 \vdash 4500$	12	24,5	91,8
$4500 \vdash 6000$	2	4,1	95,9
$6000 \vdash 7500$	2	4,1	100,0
Total	49	100,0	100,0

Obs: um dos participantes não informou o salário.

Distribuição de Frequência

- A partir dessa última tabela, é possível construir um gráfico de frequências, que recebe o nome de **histograma**.



Histograma

- O histograma é um gráfico construído a partir da distribuição de frequências agrupado por classes.

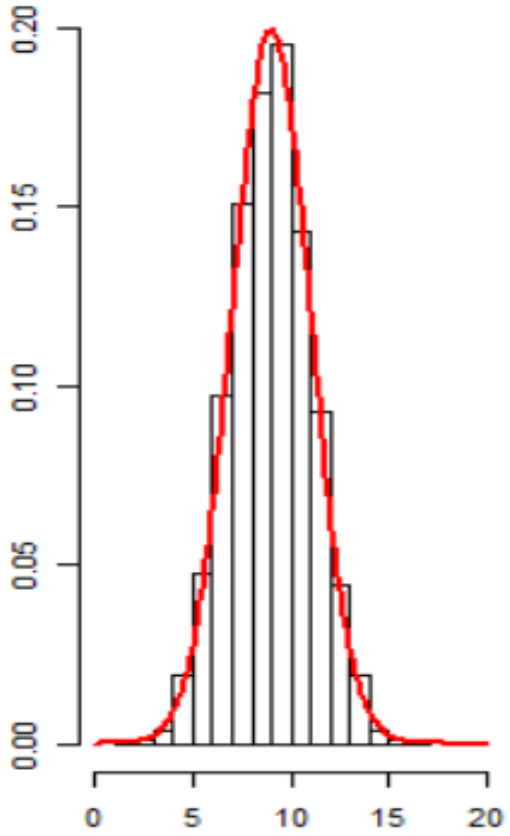
- É muito útil para observar uma grande quantidade de dados (n > 30) de forma agrupada.

- A partir do histograma, podemos observar os seguintes resultados:

  - ✓ Quantas vezes ocorre um certo resultado
  - ✓ Simetria ou assimentria de dados
  - ✓ Onde se concentram a maioria dos valores
  - ✓ Qual a dispersão dos dados
  - ✓ Existência de valores aberrantes (“dados suspeitos” ou “outliers”)

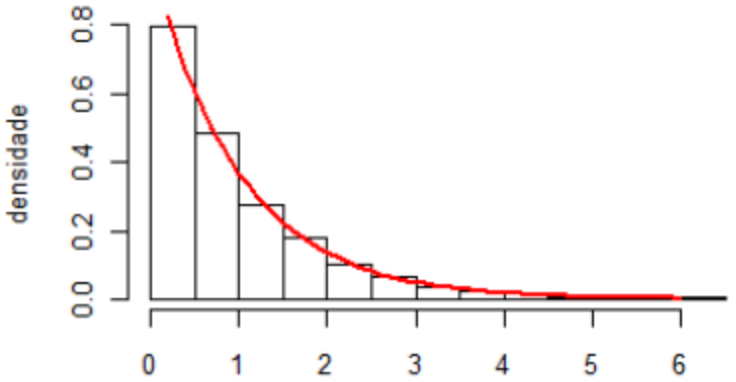
Histograma

- A seguir temos alguns exemplos de histogramas, que representam distribuições de frequências comumente encontradas:

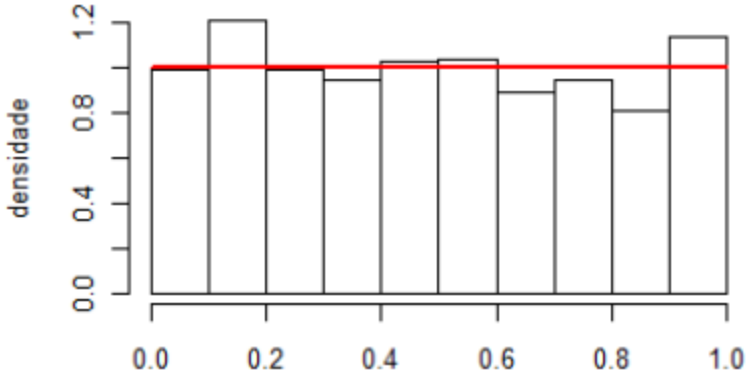


Distribuição normal (ou Gaussiana)

Distribuição exponencial



Distribuição uniforme



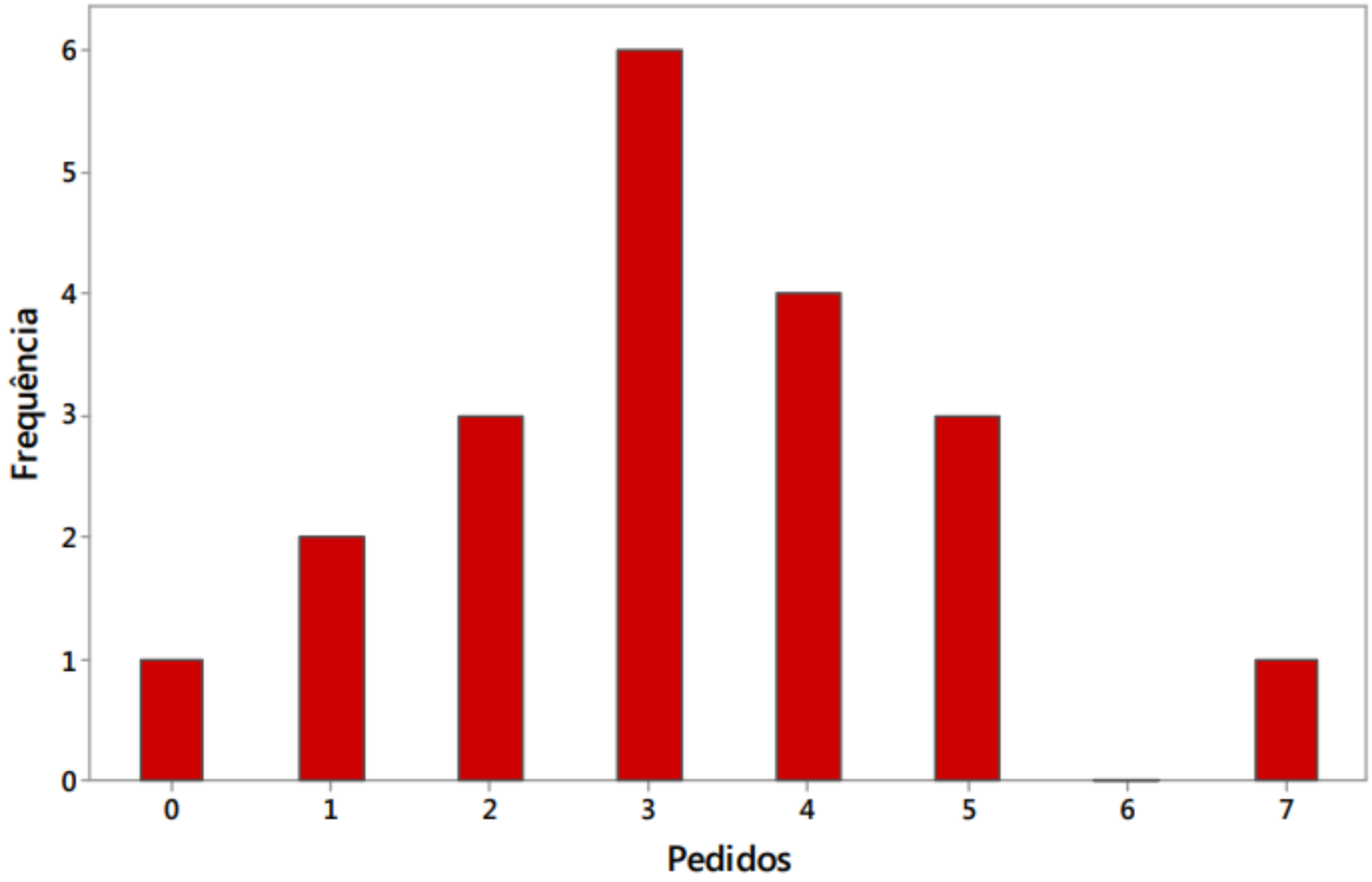
Histograma

- **Ex:** monte o histograma dos dados abaixo que representam o número de pedidos de concessão de empréstimo recebidos por uma agência nas últimas 20 semanas. OBS: utilize um tamanho de classe igual a 1.

(1) 2	(2) 3	(3) 5	(4) 4
(5) 7	(6) 4	(7) 2	(8) 5
(9) 1	(10) 3	(11) 3	(12) 5
(13) 3	(14) 4	(15) 0	(16) 3
(17) 4	(18) 1	(19) 2	(20) 3

Histograma

- **Ex:** monte o histograma dos dados abaixo que representam o número de pedidos de concessão de empréstimo recebidos por uma agência nas últimas 20 semanas. OBS: utilize um tamanho de classe igual a 1.



Histograma

- Como definir o tamanho de cada classe de dados?

- Não existem regras fixas para definir o tamanho da classe de um histograma, entre os diversos métodos disponíveis, temos os seguintes que se baseiam no número de classes:

  - **Raiz-quadrada da amostra:** $k = \sqrt{N}$

  - **Fórmula de Sturges:** $k = 1 + 3{,}32 * \log_{10}(N)$

OBS: o número de classes é um número inteiro, portanto deve ser arredondado para o valor mais próximo

Histograma

- **Exemplo:** uma pesquisa sobre economia de água quantificou o tempo médio (em minutos) gasto durante o banho de 30 pessoas. Monte uma tabela de distribuição de frequências e o histograma da distribuição e determine o percentual de pessoas que gastam menos do que 18 minutos no banho. **OBS:** os dados da tabela já estão ordenados.

2	3	4	5	5	5	5	6	7	8
8	8	9	10	10	12	12	14	14	14
16	20	23	25	25	28	30	32	35	38

Histograma

- **Exemplo 2:** Apenas a partir da tabela de distribuição criada a partir dos dados do exemplo anterior, determine o Q1, Q2 e o Q3 para desenhar o boxplot da distribuição.

- **Exemplo 2.1:** Utilizando os dados do exemplo 1, implemente em pyhon um programa que gere o histograma da distribuição.

- Aqui podemos utilizar as seguintes funções do `matplotlib` e do `numpy`:
  - `histogram(dados, bins=10)`: produz a tabela de distribuições do histograma. Usa 10 intervalos como padrão. Pode também usar uma lista para pré-definir os intervalos.
  - `hist(dados, bins=10)`: plot o histograma dos dados. Usa a função anterior como base.
  - `hist(bins[:-1], bins, weights=freq)`: outra forma de utilizar a função hist, onde já temos os intervalos e as frequências.

Histograma

Exemplo 3: desenhe o histograma dos dados abaixo, que representam o percentual recolhido de imposto de 50 diferentes empresas:

Souza Cruz	36,9
Autolatina	28,4
General Motors	25,7
Brahma	65,4
Philip Morris	46,7
Shell	3,4
Gessy Lever	15,8
IBM	20,6
Fiat Automóveis	14,2
Nestlé	9,0
Goodyear	15,1
Esso	1,8
Mercedes-Benz	1,9
Firestone	32,7
Pirelli	17,6
Texaco	3,8
Atlantic	4,5
Skol	37,5
Consul	14,3
Santa Marina	20,3
CBA	7,7
Antarctica Paulista	46,5
Brastemp	12,1
Suzano	12,7
Philips	4,8

Petróleo Ipiranga	3,1
Johnson & Johnson	10,1
Avon	17,8
Antarctica – Rio	28,9
Alcan	7,8
Bosch	18,9
Klabin	11,0
Glasurit	11,1
Kaiser – SP	56,1
Krupp	27,0
Carrefour	2,4
Usiminas	5,2
3M	26,0
Hoechst	8,0
Poliolefinas	22,9
Cebrasp	30,0
Arno	13,9
MBR	8,0
Estrela	3,4
Solvay	13,3
Kodak	10,2
Metal Leve	14,2
Champion	9,1
Rhodia	4,8
Antarctica – Nordeste	29,4