

Análise Estatística de dados

Inteligência Artificial



AULA 06 – NOÇÕES BÁSICAS DE ESTATÍSTICA

Arturo Forner-Cordero
Larissa Driemeier

PROGRAMA DO CURSO

Aula	Data	Conteúdo da Aula
01	27/02	Aula Inaugural
02	05/03	Introdução ao Curso. Noções de Álgebra Linear , Geometria Analítica Parte I
03	12/03	Noções de Álgebra Linear , Geometria Analítica Parte II
04	19/03	Decomposição de valor singular (SVD)
05	26/03	Otimização: derivadas, derivadas parciais (operadores gradiente, Jacobiano, Hessiano e Laplaciano), algoritmos de gradiente
06	02/04	Variáveis independentes e não independentes. Estatística Descritiva e Indutiva. Definições de medidas de dispersão e tendência central.
07	09/04	Probabilidade e Teorema de Bayes
08	16/04	Modelos de probabilidade discretos
09	23/04	Modelos de probabilidade contínuos
10	30/04	Modelo de Markov. Modelo de Markov oculto.

A ESTATÍSTICA

9	5	1	8		2	
3		6		9	1	4
4	2		3		7	5
3			1	6	8	
		4	7			
2	9		8		4	
6	9		8	5	3	
5	3	7	4			
7			6	8	9	

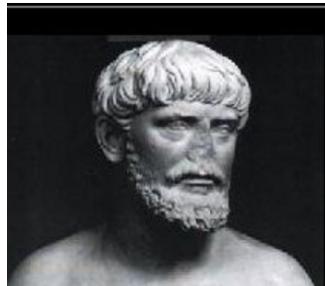
Exemplos extraídos do site:
<http://www.pagefiller.com/FeatureGallery.aspx>

20	5		16		12	16	13	23
	10	13						
			10		19			
23		9	4					
			14			7	21	
18	21			15				
	14	16		4		21	8	15
			13					
				17		8		

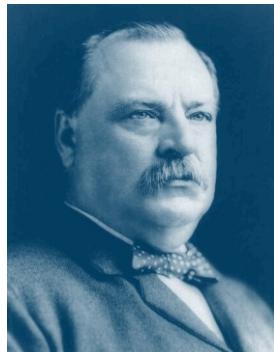


Introdução

PORQUE NÃO DEVEMOS TENTAR PREVER O FUTURO?



“As invenções há muito atingiram seu limite - e não vejo esperança para novos desenvolvimentos.”, Julius Frontinus, famoso engenheiro e senador romano”, 10 AD.

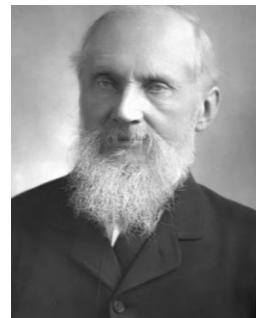


“Mulheres sensíveis e responsáveis não gostariam de votar.” Grover Cleveland, presidente americano, 1905.

“Os americanos precisam de telefones, os ingleses não. Nós temos um grande número de mensageiros em nosso país”, Willian Preece, engenheiro chefe do Escritório Geral dos Correios da Inglaterra, 1876.



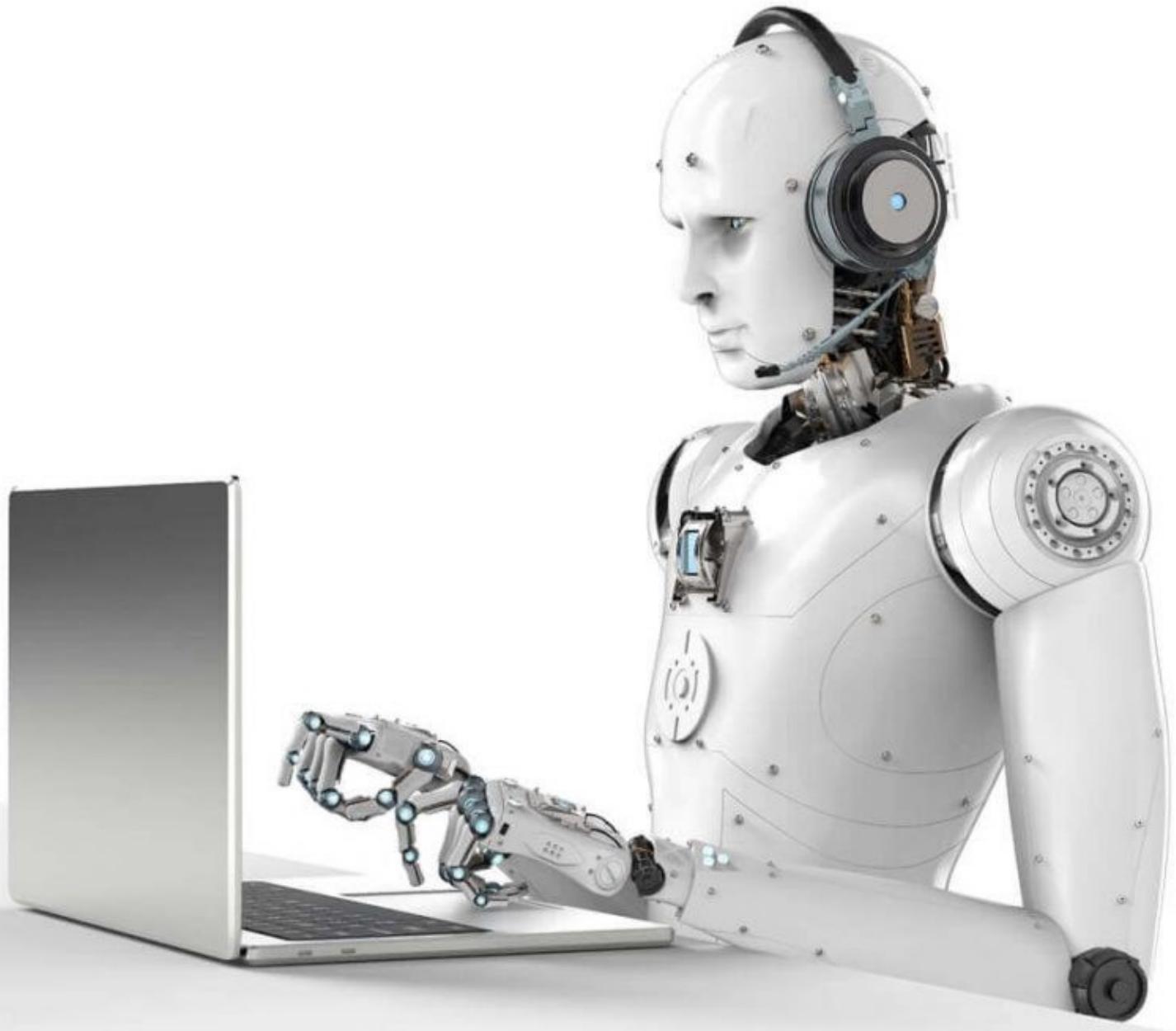
“A abolição da dor na cirurgia é uma quimera. É um absurdo continuar procurando. . . . Faca e dor são duas palavras em cirurgia que devem estar associadas para sempre na consciência do paciente.” Dr Alfred Velpeau, cirurgião francês, 1839.



“Máquinas de voar mais pesadas do que o ar são impossíveis.” Lord Kelvin, 1895.



“640 KB serão suficientes para qualquer um no futuro.” Bill Gates , 1981. **“Daqui a dois anos, o spam estará resolvido.”** Bill Gates, 2004.



QUEM NUNCA OUVIU???

O otimismo é uma estratégia para fazer um futuro melhor. Porque, a menos que você acredite que o futuro pode ser melhor, é improvável que você intensifique e assuma a responsabilidade por fazê-lo. Se você assumir que não há esperança, garante que não haverá esperança.

Avram Noam Chomsky



AI está aqui e vai mudar tudo! OMG o céu está caindo! Os programadores agora estão obsoletos. Não, eles são necessários mais do que nunca. Grandes modelos de linguagem destruirão o jornalismo, a democracia, a sociedade. Não, eles nos livrarão do trabalho penoso e seremos todos mais felizes. O câncer será resolvido por causa da IA. As alucinações da IA darão início a uma nova era de desinformação.

PORQUE ESTUDAR ESTATÍSTICA?

Você pode ser otimista. Na vida nada é certo.

**Para tudo que fazemos, avaliamos a chance
de sermos bem-sucedidos!**

**Em todos os níveis, alguém está processando
números e usando dados para orientar sua
tomada de decisão.**

2005 – 130 EXABYTES

2010 – 1,200 EXABYTES

2015 – 7,900 EXABYTES

2020 – 40,900 EXABYTES

In 2010 Eric Schmidt, Google CEO, estimou "*There was 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing.*"

Robert J. Moore, fundador e CEO da RJMetrics, em 2011, corrigiu afirmando que "*23 Exabytes of information was recorded and replicated in 2002. We now record and transfer that much information every 7 days.*"

Embora os críticos afirmem que ninguém sabe ao certo quanta informação foi produzida ou quanto rápido foi o aumento da produção, certamente estamos produzindo muita informação e em velocidades cada vez maiores.

DATA IS THE NEW OIL

Em 2006, Humby cunhou a frase
“Data is the new oil”



DATA IS THE NEW OIL

Em 2006, Humby cunhou a frase
“Data is the new oil”

Michael Palmer expandiu a citação de Humby dizendo que, como o petróleo, os dados são **“valuable, but if unrefined it cannot really be used. Oil has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so, data must be broken down and analysed for it to have value.”**



"Tenho amigos que sempre se encontram para observar passarinhos nos Estados Unidos. À noite, as conversas são sempre sobre a beleza do canto dos pássaros. Fiquei uns cinco anos sem fazer esse programa. Fui recentemente com eles e as conversas foram sobre os dados estatísticos da migração dos pássaros. Até olhar passarinho agora se baseia em dados"

Jorge Paulo Lemann



TEMOS MUITAS PERGUNTAS

Os primeiros bebês nascem mais cedo?

O que dá mais dinheiro: um filme livre ou com censura 18 anos?

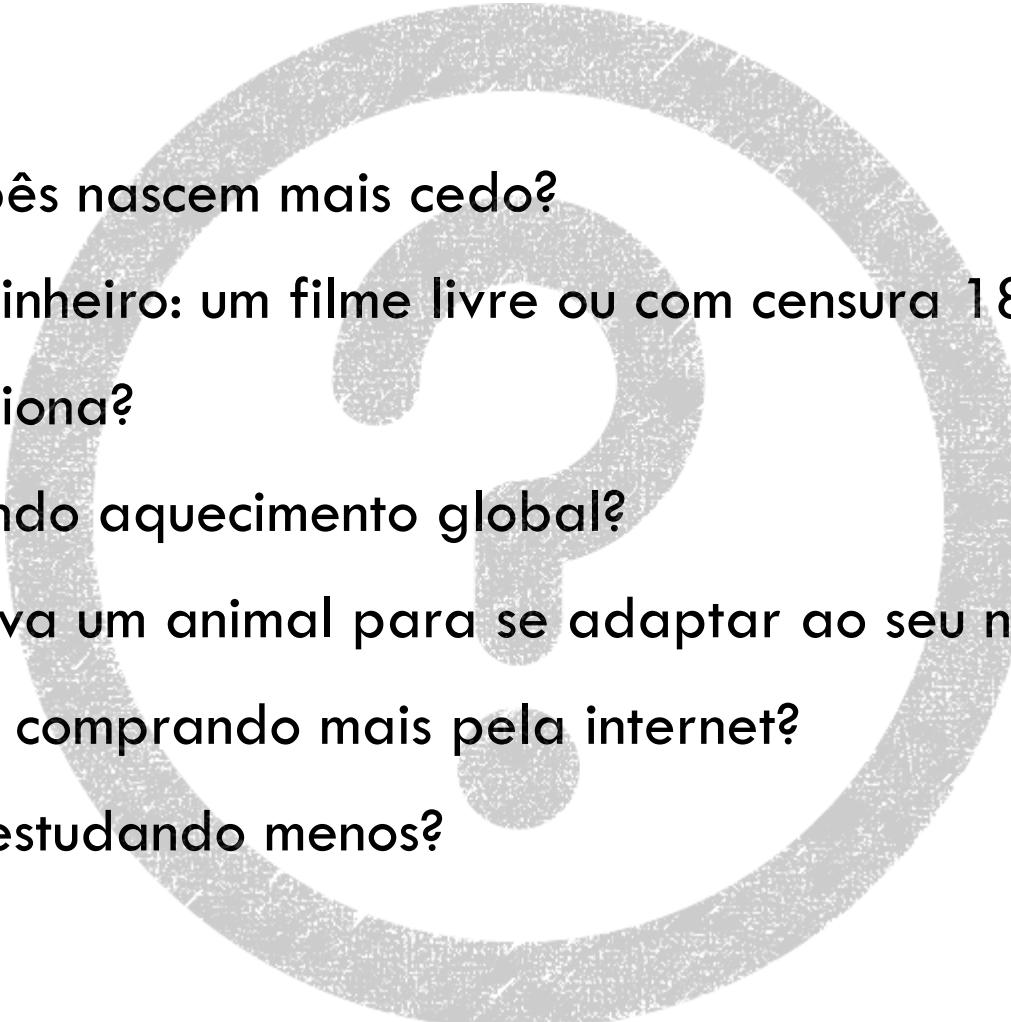
Homeopatia funciona?

Estamos vivenciando aquecimento global?

Quanto tempo leva um animal para se adaptar ao seu novo lar?

As pessoas estão comprando mais pela internet?

Os jovens estão estudando menos?



EVIDÊNCIA ANEDÓTICA

"Não é verdade que o primeiro filho nasce mais cedo... O meu primeiro chegou duas semanas atrasado e agora acho que o segundo vai sair duas semanas mais cedo!! "

"Filme livre dá muito mais dinheiro, lá em casa, só assistimos filmes censura livre para meu irmão participar..."

"Claro que homeopatia funciona, eu mesmo fui curado de uma sinusite por tratamento homeopático."

"Essa história de aquecimento global é bobagem, hoje foi um dia frio e estamos em pleno verão!"



Evidência, geralmente pessoal, que é coletada casualmente, e não por um estudo bem planejado.

ACHISMOS, SENSO COMUM E EVIDÊNCIAS ANEDÓTICAS.

Número pequeno de observações: alguns poucos casos, geralmente difíceis de serem comprovados, não têm relevância estatística para suportar uma tese defendida;

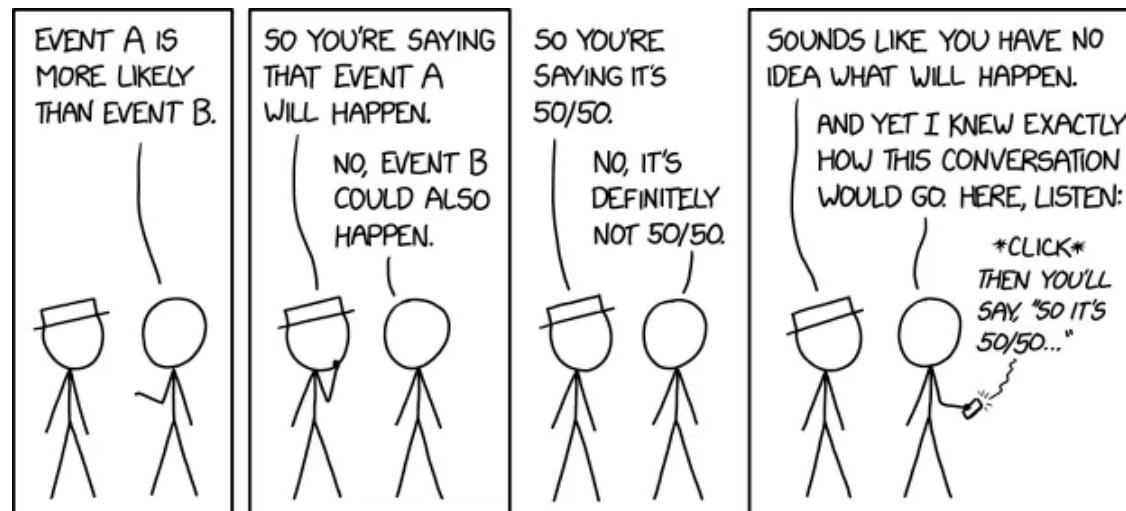
Viés de confirmação: fenômeno cognitivo o qual consiste na tendência de uma pessoa em reparar e lembrar com facilidade as evidências favoráveis às suas crenças, ignorar e esquecer evidências desfavoráveis, ou ainda interpretar evidência ambígua como evidência favorável;

Imprecisão: as anedotas são frequentemente histórias pessoais e, muitas vezes, esquecidas, deturpadas, repetidas incorretamente, etc...

O QUE É PENSAMENTO DETERMINÍSTICO?

Simplificando, o determinismo é “se x , então y ”. É a noção de que se fizermos algo (x), um determinado resultado (y) é inevitável.

Os determinismos não são inherentemente positivos ou negativos. É tão determinista dizer “se construirmos a mídia social, o mundo será um lugar mais conectado” quanto dizer “a mídia social destruirá a democracia”.

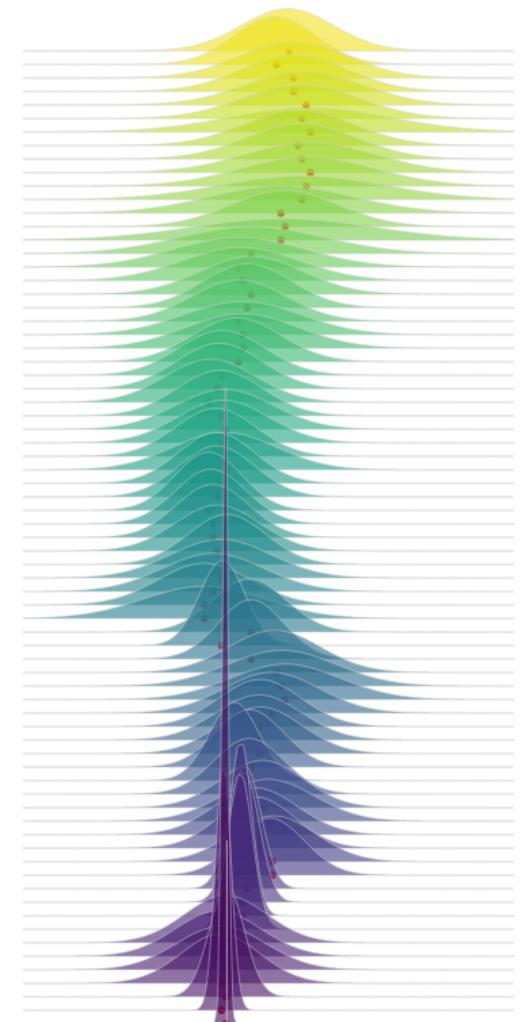


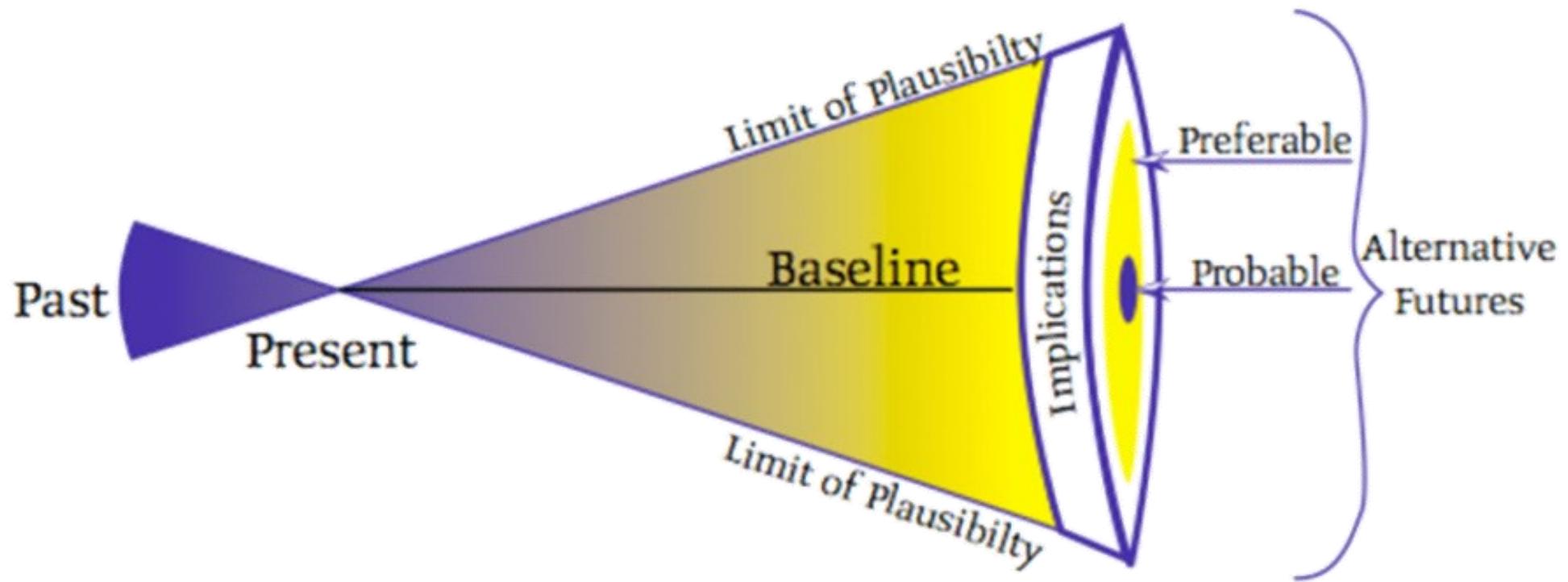
ABRAÇANDO FUTUROS PROBABILÍSTICOS

“O contraponto ao determinismo não é o indeterminismo. É insatisfatório jogar as mãos para o alto e dizer “qualquer futuro é possível”. Alguns futuros são mais propensos a ocorrer do que outros. Alguns resultados são mais prováveis devido a uma intervenção técnica específica do que outros.

A chave para entender como as tecnologias moldam o futuro é lidar de forma holística com como uma interrupção rearranja o cenário. Uma ferramenta é o pensamento probabilístico.

<https://zephoria.medium.com/resisting-deterministic-thinking-52ef8d78248c>





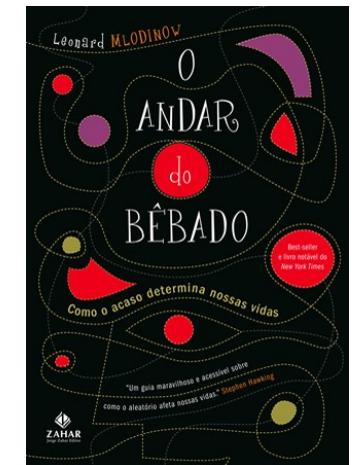
The Future is many not one

ALEATORIEDADE GERA INCERTEZAS...

“O maior desafio à compreensão do papel da aleatoriedade na vida é o fato de que, embora os princípios básicos dela surjam da lógica cotidiana, muitas das consequências que se seguem a esses princípios provam-se constraintuitivas.”

“Geralmente subestimamos o efeito da aleatoriedade”

O andar do bêbado
Leonard Mlodinow
Ed. Zahar



EXEMPLOS...

"Alguns anos atrás, um homem ganhou na loteria nacional espanhola com um bilhete que terminava com o número 48. Orgulhoso por seu feito, ele revelou a teoria que o levou à fortuna. "Sonhei com o número 7 por 7 noites consecutivas", disse, "e 7 vezes 7 é 48"

"Muitos usuários do iPod Shuffle duvidavam da aleatoriedade com que as músicas eram tocadas, pois algumas vezes um mesmo artista ou música tocava mais de uma vez. A saída encontrada por Steve Jobs, segundo o livro, foi reprogramar o iPod para que evitasse repetições e deixá-lo menos aleatório para parecer mais aleatório."

Esses parágrafos de O Andar do Bêbado de Leonard Mlodinow dão um exemplo de quão grande é a nossa capacidade de abstrair a realidade em prol da nossa percepção da realidade.

PROBABILIDADE

Probabilidade é um conceito filosófico e matemático que permite a quantificação da incerteza. Dessa maneira, ela pode ser aferida, analisada e usada para a realização de previsões ou para a orientação de intervenções.

É a Probabilidade que torna possível lidar de forma racional com problemas envolvendo o imprevisível.

PROBLEMA DA LINDA...

O clássico "problema da Linda"(Tversky & Kahneman, 1983, p. 297): "*Linda tem 31 anos, é solteira, extrovertida e muito inteligente. Ela é formada em filosofia. Como estudante, ela era bastante preocupada com questões de discriminação e justiça social, participava de manifestações e de demonstrações antinucleares.*"

Baseando-se no texto, **qual das seguintes alternativas é a mais provável:**

- Linda é caixa de banco;
- Linda é caixa de banco e ativista do movimento feminista.



AULA INTRODUTÓRIA:

Conceito de
Estatística
Probabilidade

ESTATÍSTICA E PROBABILIDADE

A probabilidade lida com a previsão de ocorrência de eventos futuros, enquanto as estatísticas envolvem a análise da frequência de eventos passados.

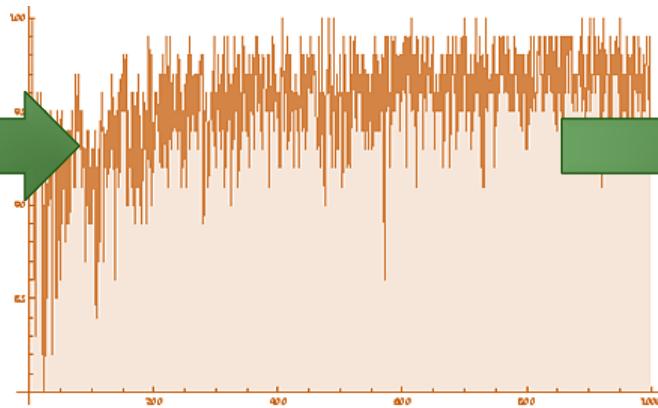
A teoria estatística e o conhecimento de probabilidade tornam o profissional crítico na análise de informações e menos sujeito a afirmações equivocadas.

ESTUDO DA ESTATÍSTICA

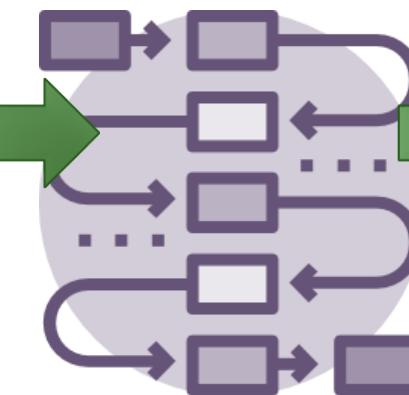
A ciência Estatística é dividida basicamente em duas partes:
Estatística Descritiva (Análise Exploratória de Dados) e **Indutiva**



Dados



Organização



Modelamento



Previsão

A **Estatística Descritiva** se preocupa com a organização, descrição, sumário e apresentação gráfica de dados. Em outras palavras, descreve quantitativamente ou resume características de uma coleção de informações.

A **Estatística Indutiva** está relacionada com a análise e interpretação dos dados e previsões futuras. É desenvolvida com base na teoria das probabilidades.

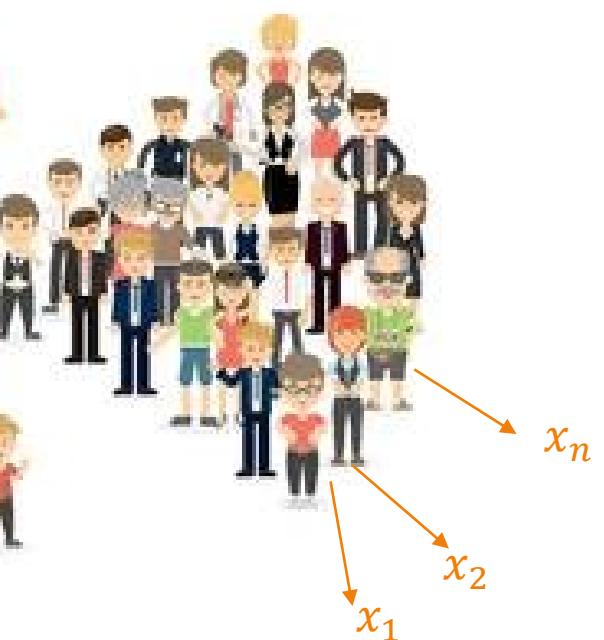
A finalidade da **Estatística Indutiva** é subsidiar conclusões sobre **populações** com base nos resultados observados em **amostras** extraídas dessas populações.

O termo “indutiva” decorre da existência de um processo de indução, em que, partindo-se do conhecimento de uma parte, procura-se tirar conclusões sobre o todo.

POPULAÇÃO



AMOSTRA



A população é a coleção completa e total dos elementos a serem considerados em um estudo estatístico

Amostra é um subconjunto de uma população de interesse

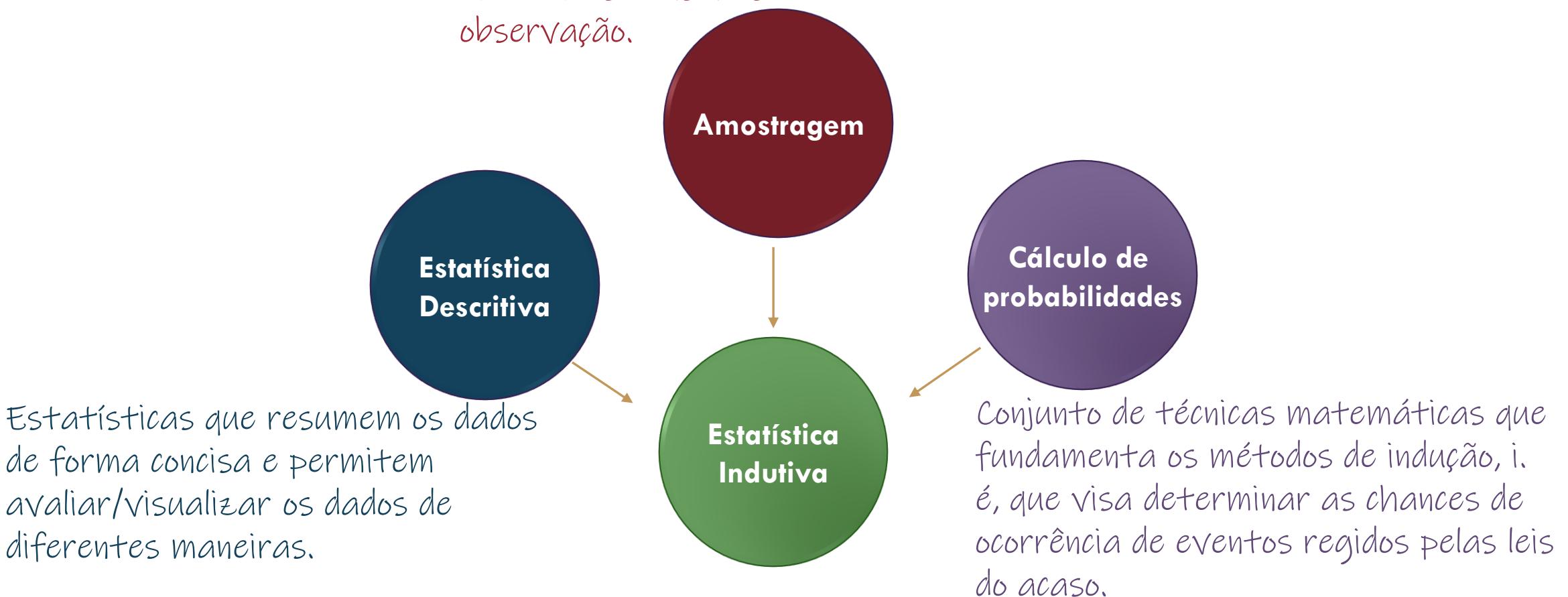
SE NOSSA CONCLUSÃO É BASEADA EM UMA AMOSTRA...

Não podemos dizer que a conclusão é falsa ou verdadeira...

... pois foram verificadas sobre um conjunto restrito de indivíduos, e portanto não são falsas, mas não foram verificadas para todos os indivíduos da População, pelo que também não podemos afirmar que são verdadeiras !

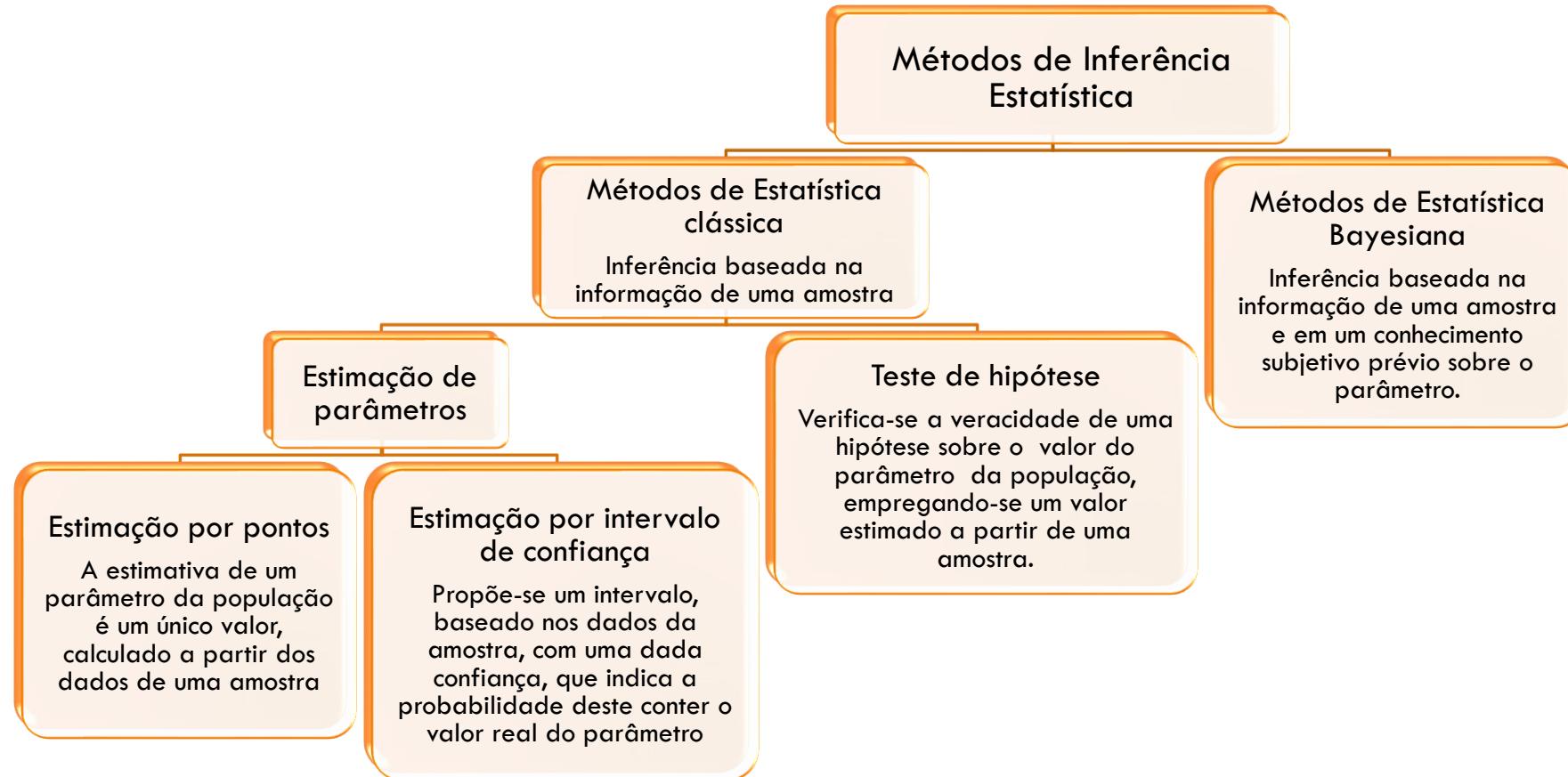
Existe, assim, um certo grau de incerteza (percentagem de erro) que é medido em termos de Probabilidade.

Um estudo estatístico completo que recorra às técnicas de **Estatística Indutiva** irá envolver também os seguintes tópicos:



Fonte: Costa Neto, P. L. "Estatística", 1^a edição, Editora Edgard Blucher, 1983.

ESTATÍSTICA INDUTIVA



Fonte:

Krishnamoorthi, K.S. *Reliability Methods for Engineers*, 1^a edição, ASQC QualityPress, 1992.

VAMOS TREINAR...

Assinale a(s) alternativa(s) correta(s).

- Um jornal americano realizou, nos EUA, uma pesquisa com 800 pessoas divorciadas perguntando se elas desejavam se casar novamente. Foi encontrado que 58% dos entrevistados não desejam um novo casamento. Os 800 divorciados constituem uma população.
- A **Estatística Descritiva** descreve quantitativamente ou resume características de uma coleção de informações.
- A **Estatística Indutiva** está relacionada com a análise e interpretação dos dados e previsões futuras.
- A **Estatística Descritiva** está baseada na teoria das probabilidades.



ESTATÍSTICA DESCRIPTIVA

Definir as categorias de dados
Definir e calcular medidas de tendência central
Definir e calcular medidas de dispersão
Construir e usar tabelas de frequências

ESTATÍSTICA DESCRIPTIVA

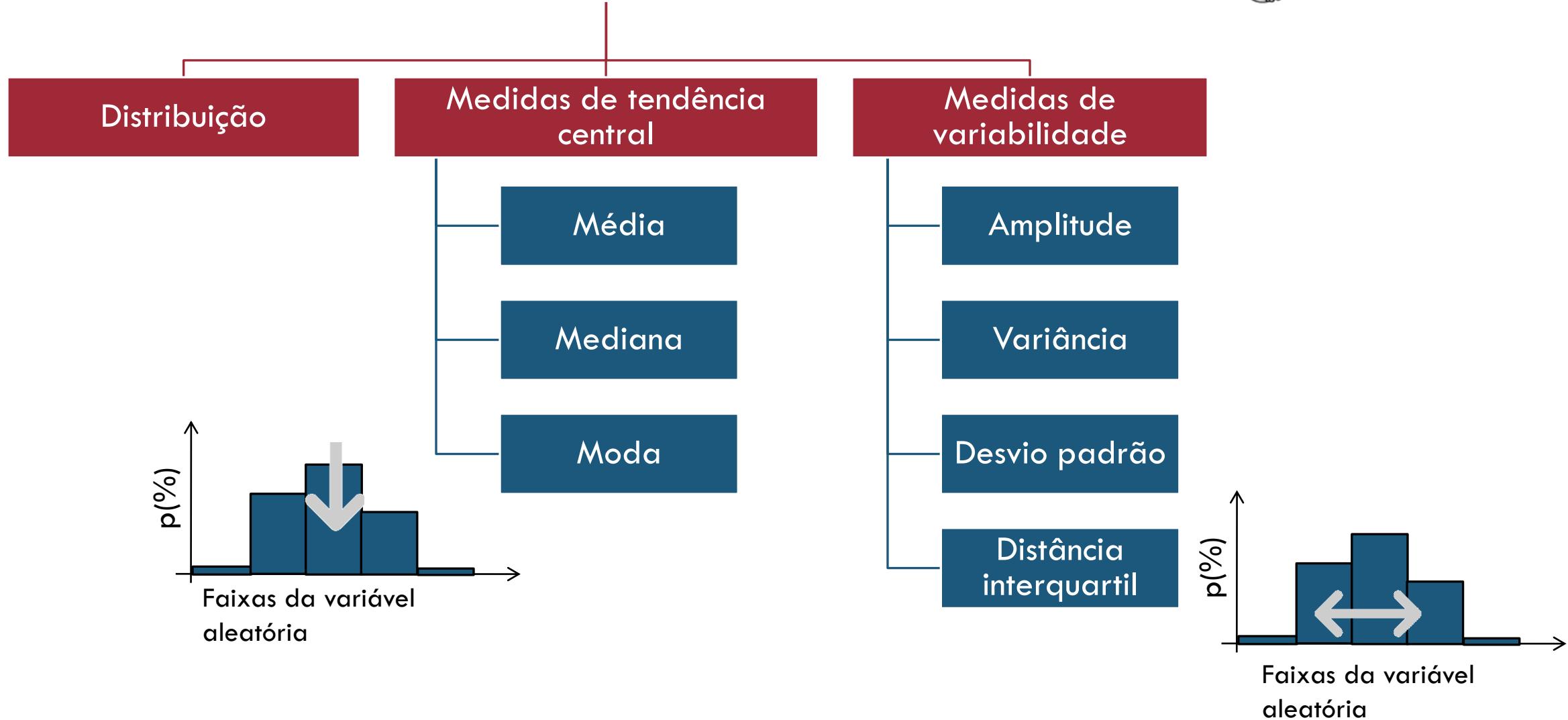
Quando se avalia uma característica de interesse de um produto ou serviço através de um conjunto de medidas, **não estamos interessados em cada medida individual**, mas sim no padrão de comportamento como um todo.

O padrão pode ser caracterizado por alguns poucos números e gráficos que quantificam e exibem informações importantes.

O sumário dos aspectos importantes de um conjunto de dados é chamado de **Estatística Descritiva**.



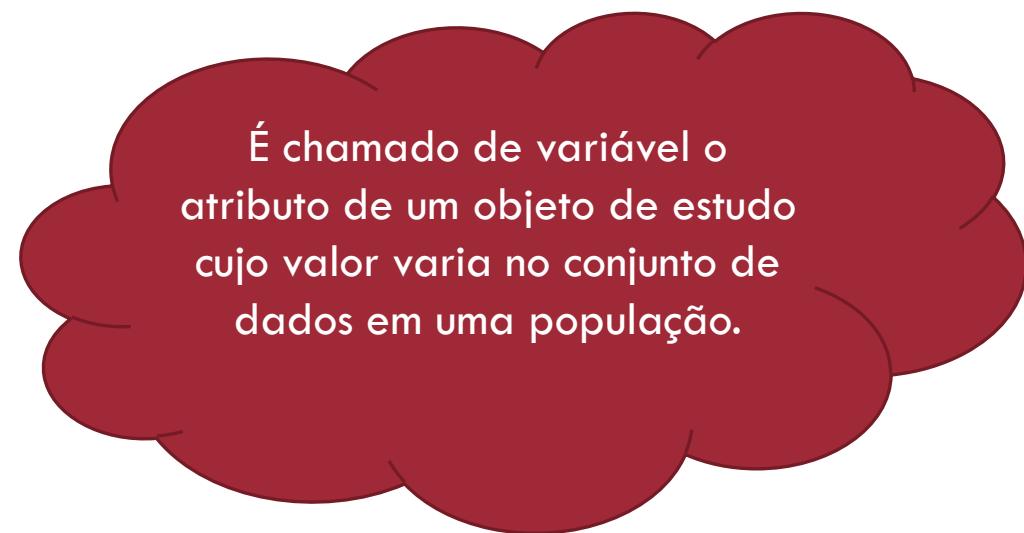
Estatística descritiva



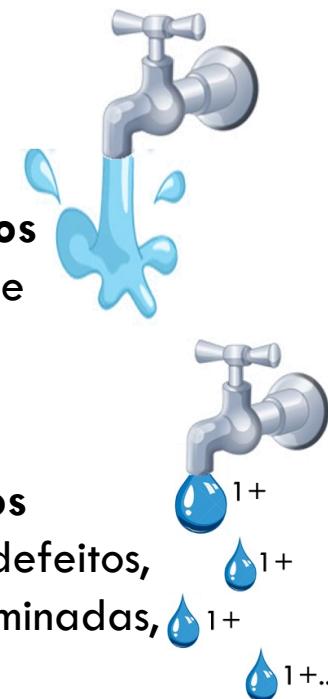
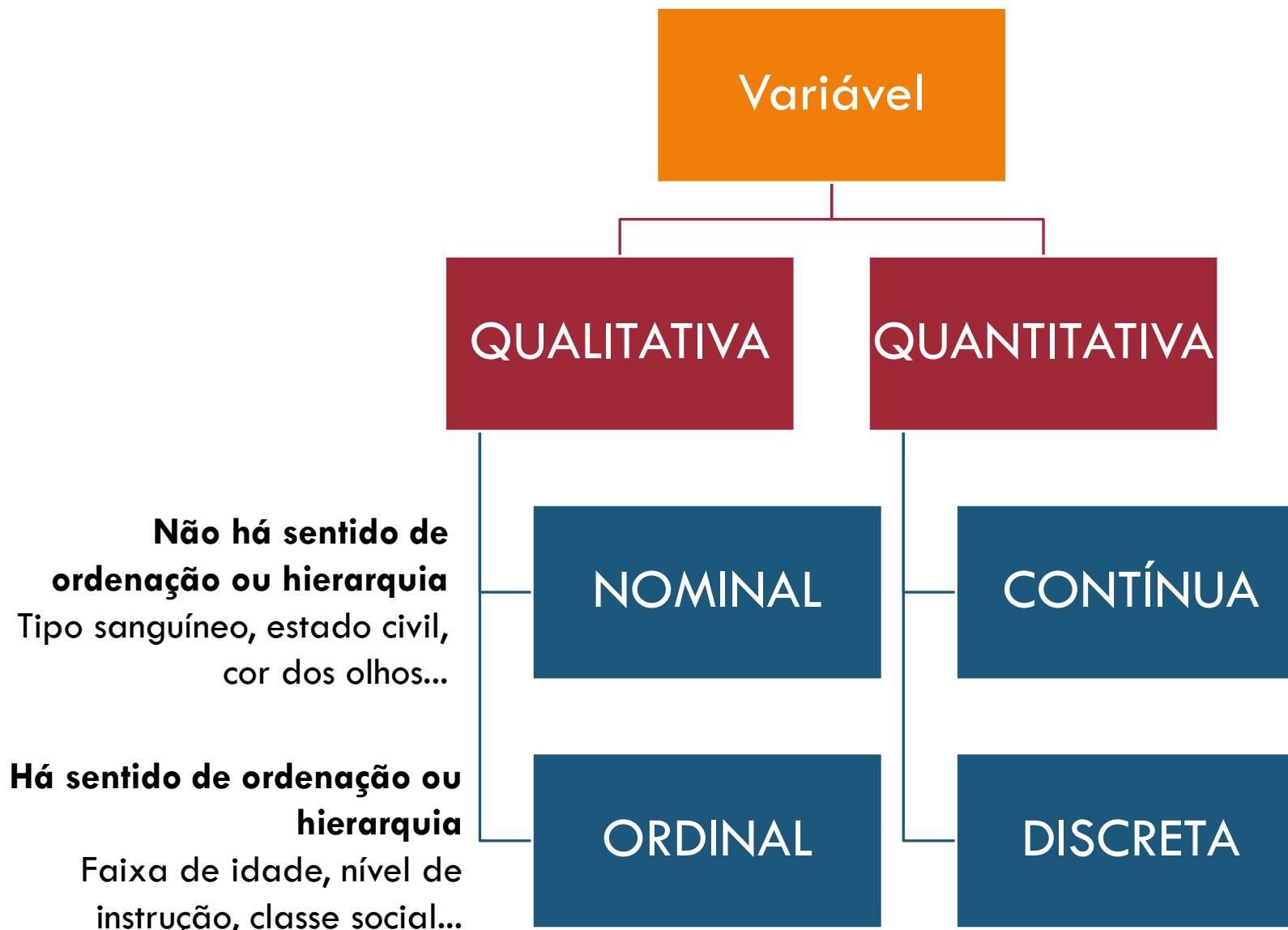
TIPOS DE DADOS

O tipo de dado que está sendo analisado é importante para ajudar a determinar o tipo de exposição visual, análise de dados ou modelo estatístico. Inclusive, softwares de ciência de dados, como R ou Python, utilizam esses tipos de dados para melhorar seu desempenho computacional.

Além disso, o tipo de dado para uma **variável** determina como o software processará os cálculos para aquela variável.



É chamado de variável o atributo de um objeto de estudo cujo valor varia no conjunto de dados em uma população.



TIPOS DE DADOS QUALITATIVOS E QUANTITATIVOS

Quantitativos: podem ser contados, medidos e expressos usando números. São subdivididos em:

- **Contínuos:** podem assumir qualquer valor em um intervalo;
- **Discretos:** podem assumir apenas valores inteiros, como contagens.

Categóricos ou qualitativos: são descritivos e conceituais. Os dados qualitativos podem ser categorizados com base em traços e características. São subdivididos em:

- **Ordinais:** dados categóricos que têm uma ordem explícita;
- **Nominais:** não têm ordem explícita;

Verifique em cada um dos itens abaixo e marque as variáveis como: qualitativa nominal (N), qualitativa ordinal (O), quantitativa discreta (D) e quantitativa contínua (C):

- Cor dos olhos de uma população;
- Temperatura de uma certa região, durante um certo período do ano;
- Nível educacional;
- Vida média de válvulas de descarga hidráulica;
- Concentração de impurezas em uma amostra de leite, em mg por litro;
- Tipo sanguíneo de uma população;
- Religião de um indivíduo;
- Tempo de voo entre duas cidades;
- Procedência de cada candidato ao vestibular da USP em certo ano;
- Quantidade de estudantes de uma disciplina;
- Diâmetro de uma bola de futebol;
- Número de parafusos refugados em determinado lote;
- Tempo de reação de um indivíduo após submetido a certo estímulo;
- Classificação de um café (bom, regular, ruim...);
- Estado civil dos alunos de Estatística Básica deste ano;
- Sexo de uma criança;
- Quantidade de cômodos de uma residência;
- Consumo de bebida em BH (bebe muito, bebe pouco, não bebe).

EXERCÍCIO

Diversas características são consideradas para o cálculo dos preços de venda e de aluguel de imóveis comerciais e residenciais. Ordene as variáveis abaixo na sequência: (1) Quantitativa Contínua, (2) Quantitativa Discreta, (3) Qualitativa Ordinal, (4) Qualitativa Nominal.

Região da cidade que se encontra o imóvel

Área do imóvel

Estado de conservação do imóvel

Número de quartos do imóvel

FREQUÊNCIA

Frequência e Frequência relativa

Frequência f_i é o número de vezes que um dado valor i foi observado.

$$\sum_{i=1}^k f_i = n$$

A relação entre a freqüência observada para um dado valor e o número total de elementos observados, n , é a frequência relativa.

$$p_i = \frac{f_i}{n}, \sum_{i=1}^k p_i = 1$$

Frequência acumulada para variáveis quantitativas (ou qualitativas ordinais),

A frequência acumulada F_i é a soma das frequências de todos os valores menores ou iguais ao valor correspondente a i

$$F_i = \sum_{j=1}^i f_j$$

A **frequência relativa acumulada** P_i é dada por

$$P_i = \frac{F_i}{n}, \quad \sum_{i=1}^k P_i = 1$$

EXEMPLO: VARIÁVEIS QUANTITATIVAS DISCRETAS

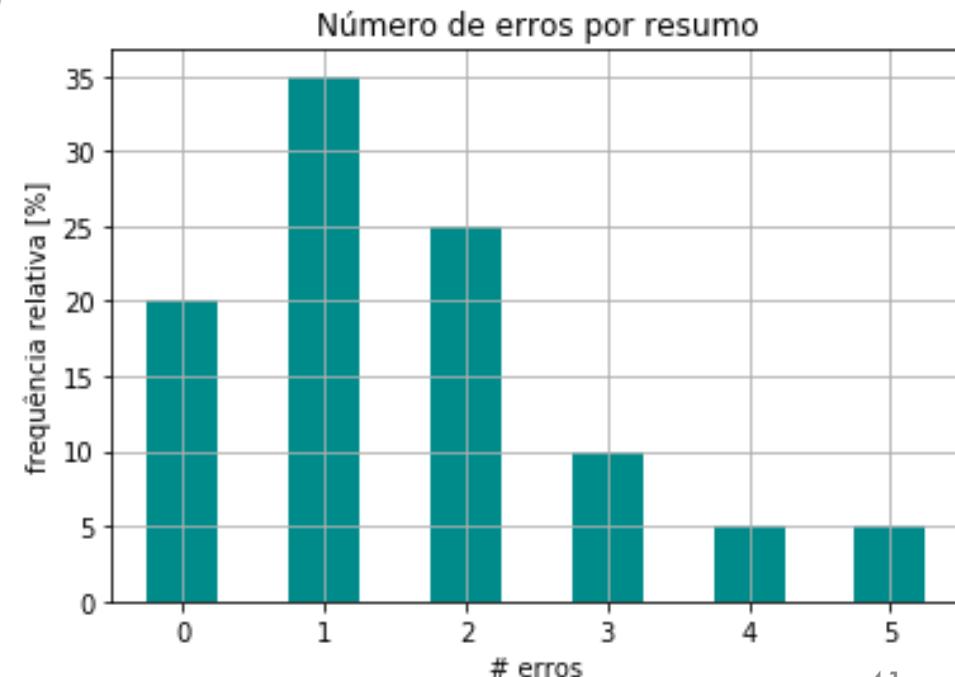
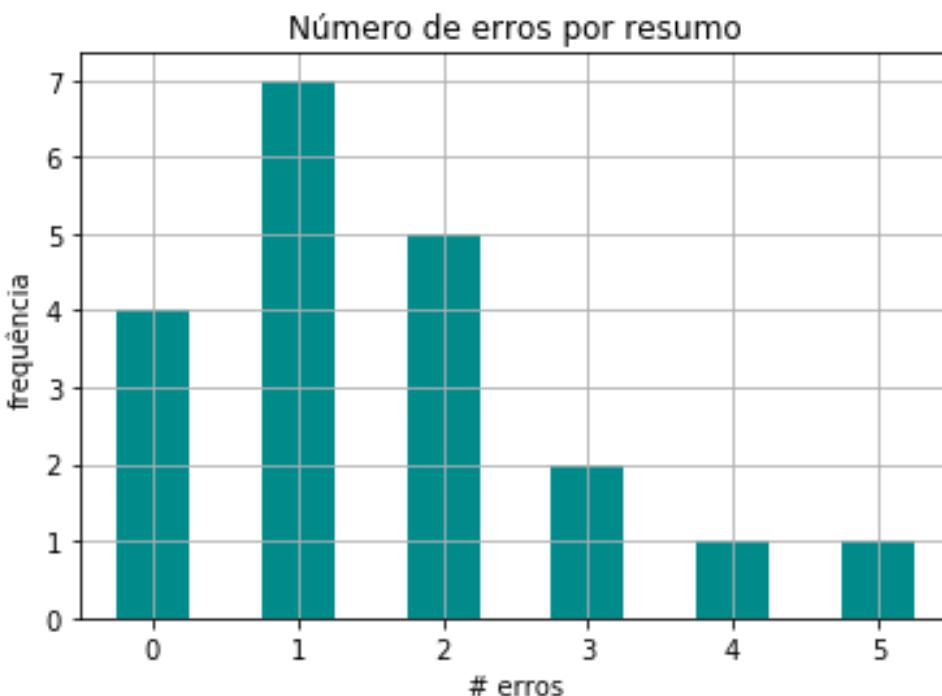
O conjunto de dados a seguir representa a variável **número de erros de inglês** observados em 20 resumos de teses extraídas do curso de Engenharia da Poli.

2	4	2	1	2
3	1	0	5	1
0	1	1	2	0
1	3	0	1	2

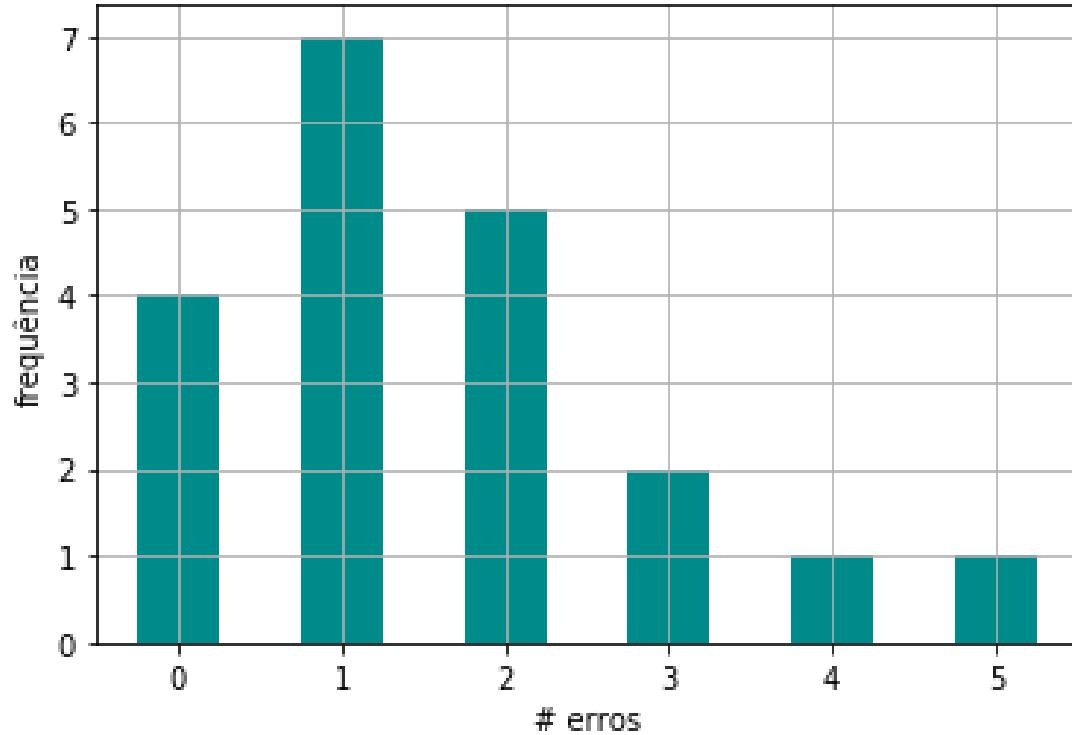
2	4	2	1
3	1	0	5
0	1	1	2
1	3	0	1

X_i	f_i	p_i	F_i	P_i
0	4	0,20	4	0,20
1	7	0,35	11	0,55
2	5	0,25	16	0,80
3	2	0,10	18	0,90
4	1	0,05	19	0,95
5	1	0,05	20	1,00
		20	1,00	

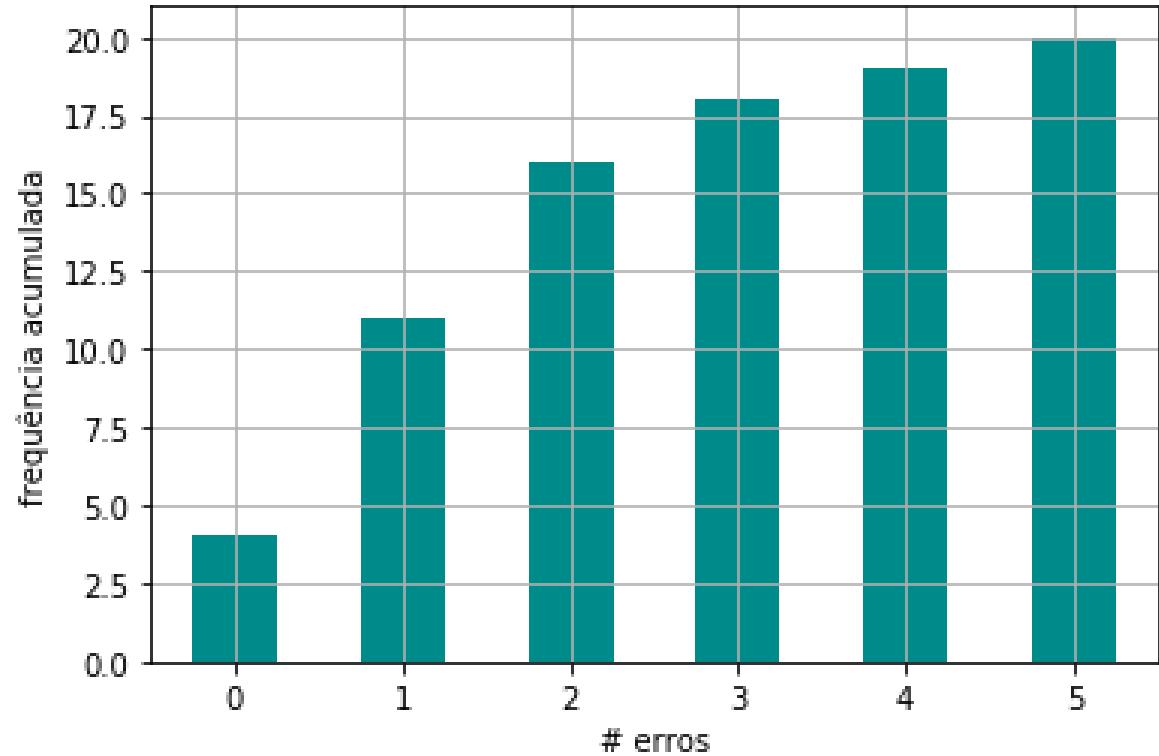
2
1
0
2



Número de erros por resumo

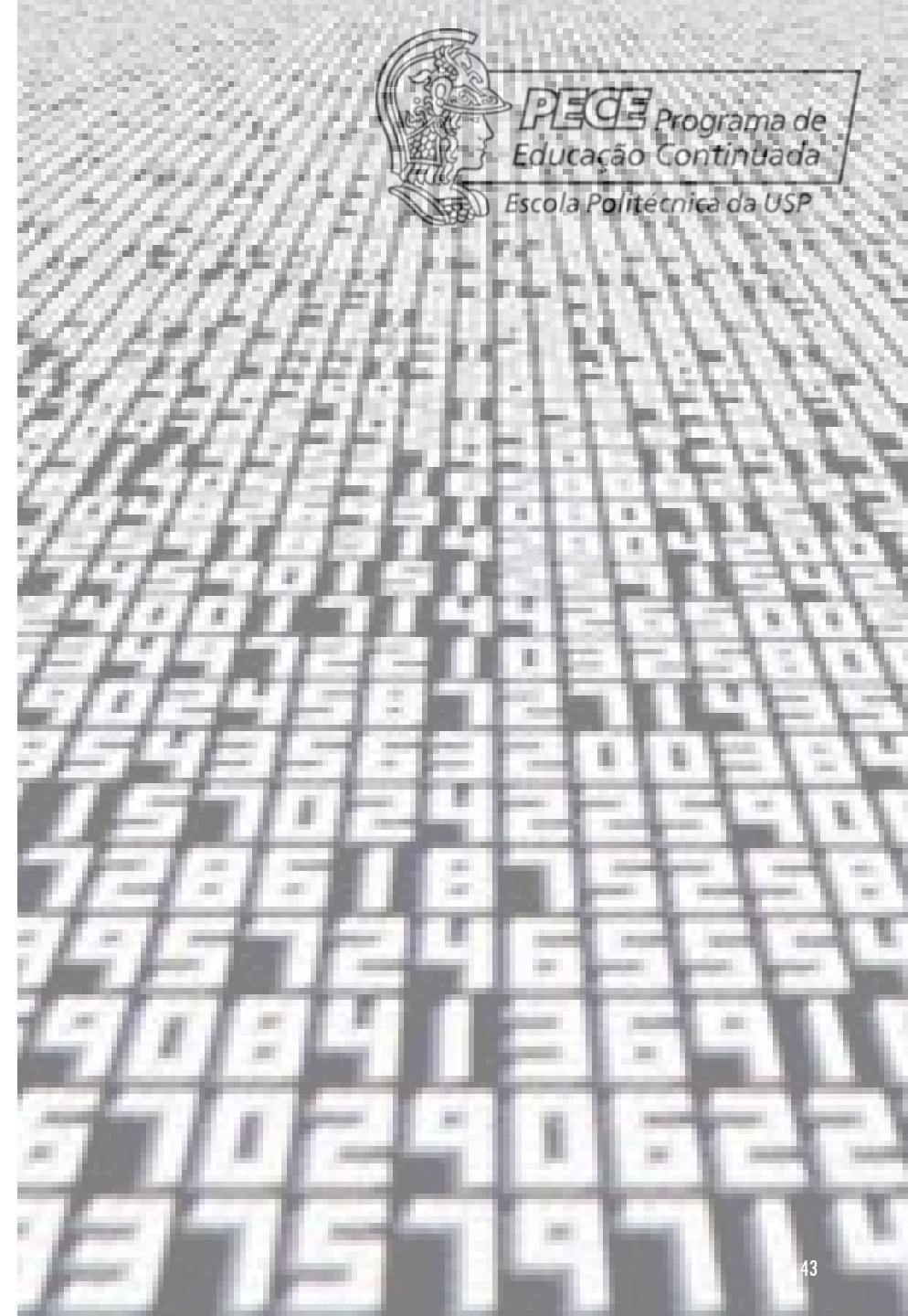


Frequência acumulada



REPRESENTAÇÃO GRÁFICA DE VARIÁVEIS

A vantagem da representação gráfica está em possibilitar uma rápida impressão visual de como se distribuem as freqüências ou as freqüências relativas no conjunto de elementos examinados.



PECe Programa de
Educação Continuada
Escola Politécnica da USP

EXEMPLO: VARIÁVEIS QUALITATIVAS NOMINAIS

Seja na tabela abaixo o número de medalhas de ouro dos 5 países mais bem colocados nos jogos paraolímpicos do Rio de Janeiro, em 2016.

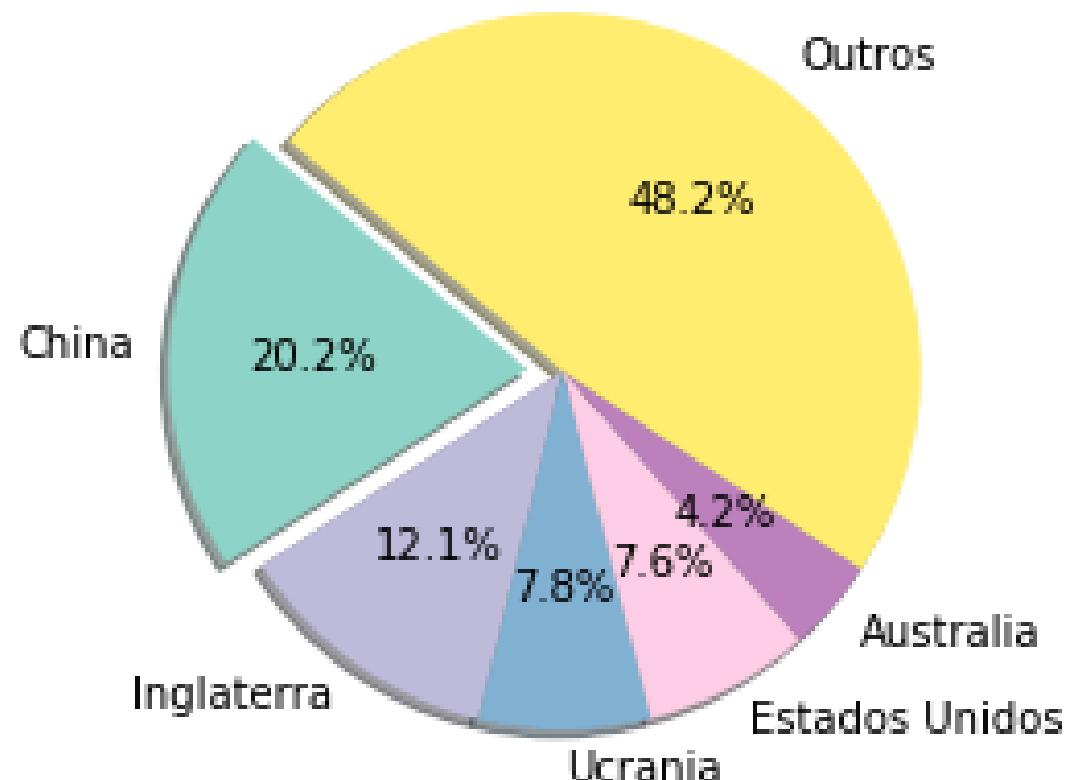
PAÍS	TOTAL	%
China	239	14.97
Inglaterra	147	9.20
Ucrânia	117	7.33
Estados Unidos	115	7.20
Austrália	81	5.07
Outros	898	56.23

Peraí!!! Isso é
variável qualitativa
nominal????



DIAGRAMA CIRCULAR

Medalhas de ouro dos cinco países mais bem colocados
Jogos paraolímpicos de 2016



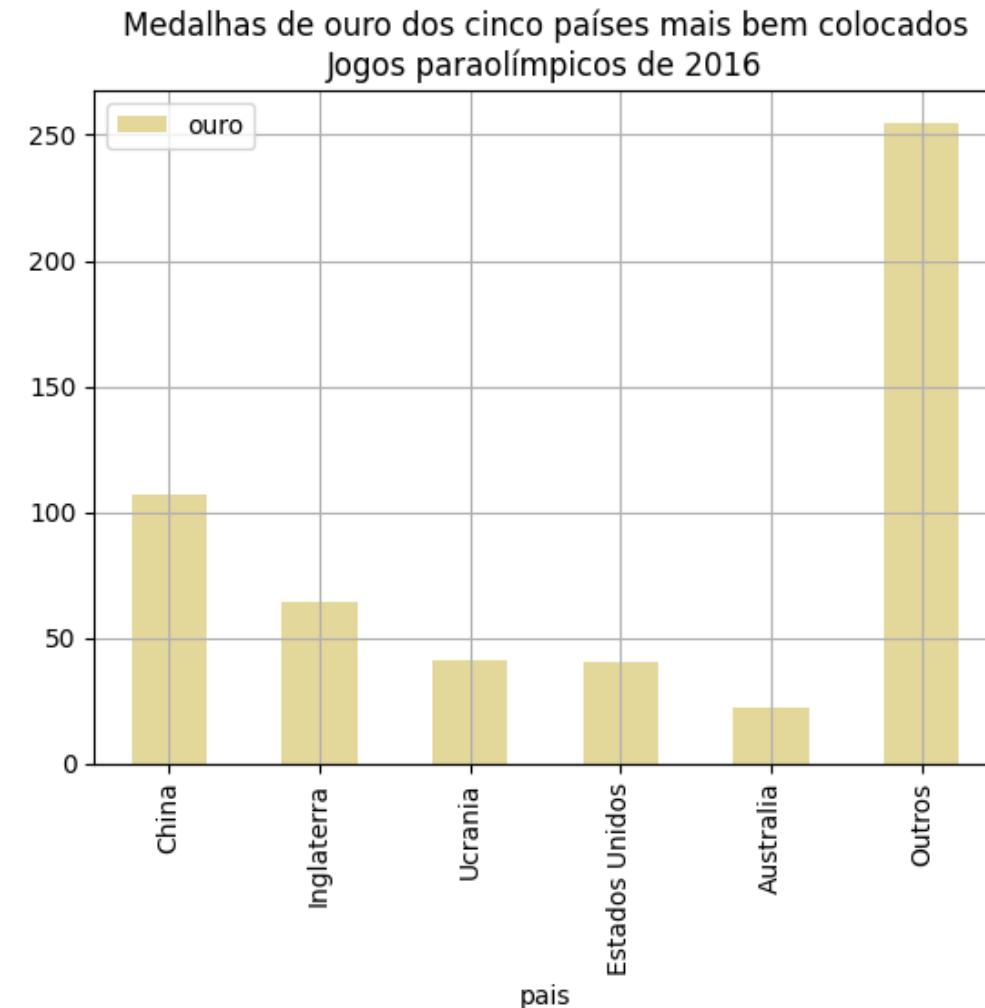
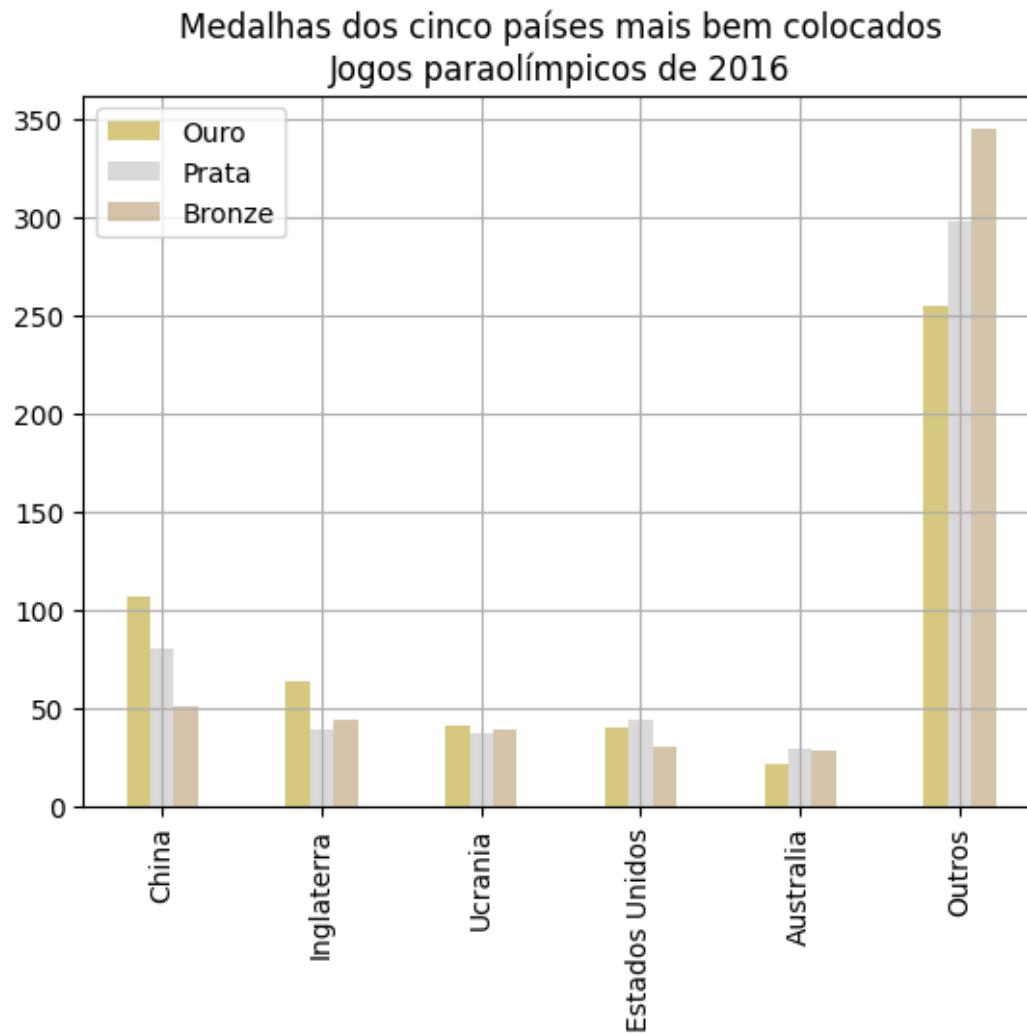
Bibliotecas python:
panda, matplotlib e pyplot

EXEMPLO: VARIÁVEIS QUALITATIVAS NOMINAIS

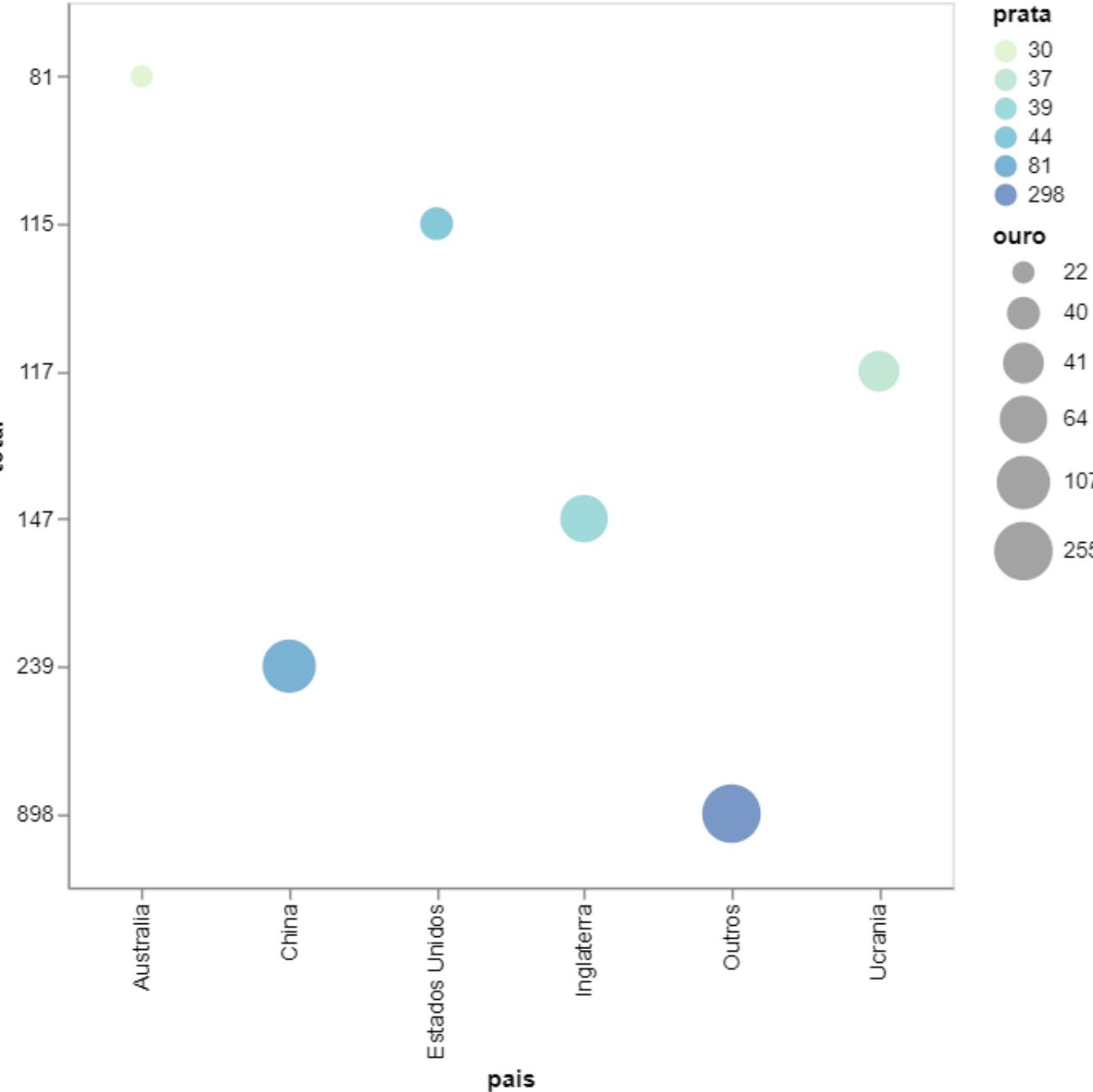
Seja na tabela abaixo o número de medalhas de ouro dos 5 países mais bem colocados nos jogos paraolímpicos do Rio de Janeiro, em 2016.

PAÍS	OURO	PRATA	BRONZE	TOTAL	%
China	107	81	51	239	14.97
Inglaterra	64	39	44	147	9.20
Ucrânia	41	37	39	117	7.33
Estados Unidos	40	44	31	115	7.20
Austrália	22	30	29	81	5.07
Outros	255	298	345	898	56.23

DIAGRAMA DE BARRAS



GRÁFICOS PODEM CONTER QUANTAS INFORMAÇÕES CONSIDERARMOS NECESSÁRIAS...^{total}



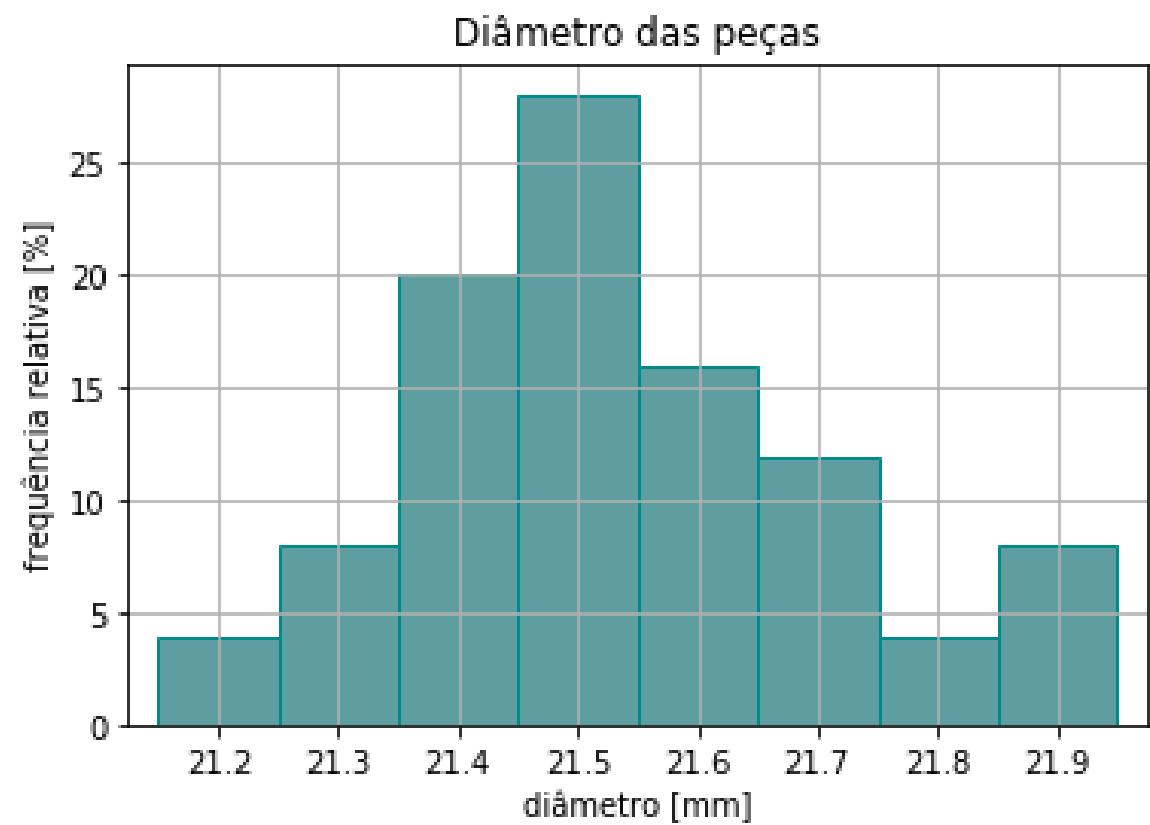
E AS VARIÁVEIS CONTÍNUAS?

Seja uma amostra de 25 valores da variável “diâmetro de peças produzidas por uma máquina”, em mm. Execute a representação gráfica.

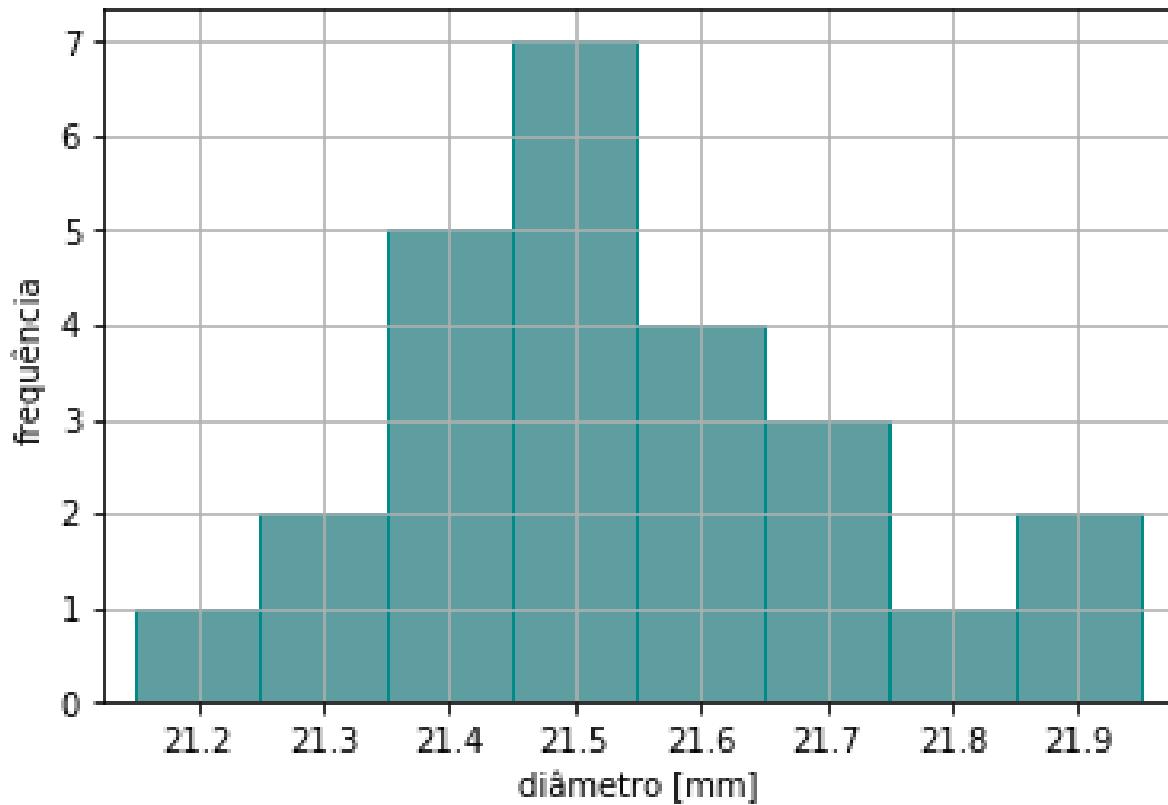
21,5	21,4	21,8	21,5	21,6
21,7	21,6	21,4	21,2	21,7
21,3	21,5	21,7	21,4	21,4
21,5	21,9	21,6	21,3	21,5
21,4	21,5	21,6	21,9	21,5

Fonte: Costa Neto, P. L. “Estatística”, 1º edição,
Editora Edgard Blucher, 1983.

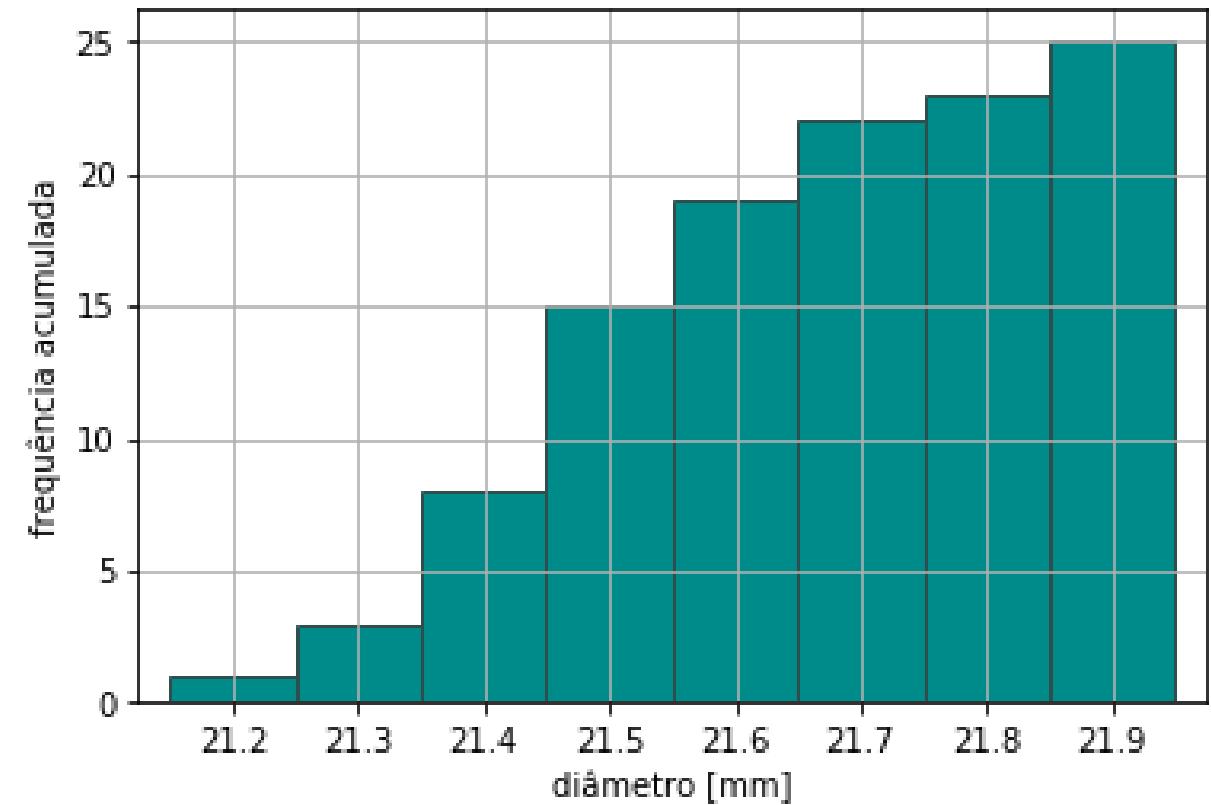
X_i	f_i	F_i	p_i	P_i
21,2	1	1	0,04	0,04
21,3	2	3	0,08	0,12
21,4	5	8	0,20	0,32
21,5	7	15	0,28	0,60
21,6	4	19	0,16	0,76
21,7	3	22	0,12	0,88
21,8	1	23	0,04	0,92
21,9	2	25	0,08	1,00
	25		1,00	

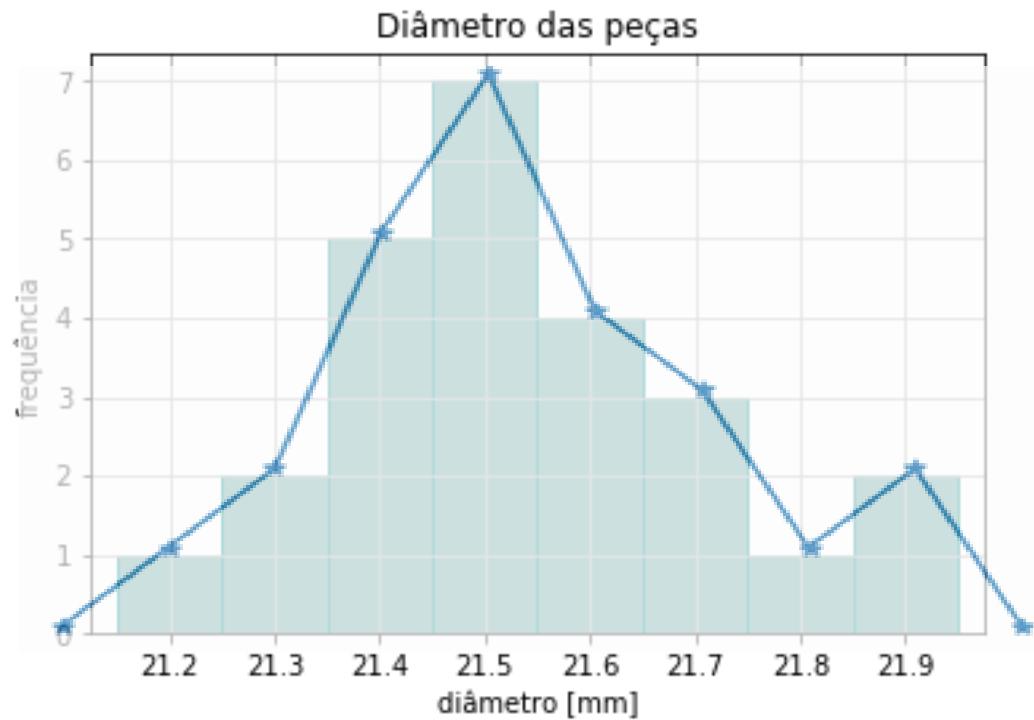


Diâmetro das peças

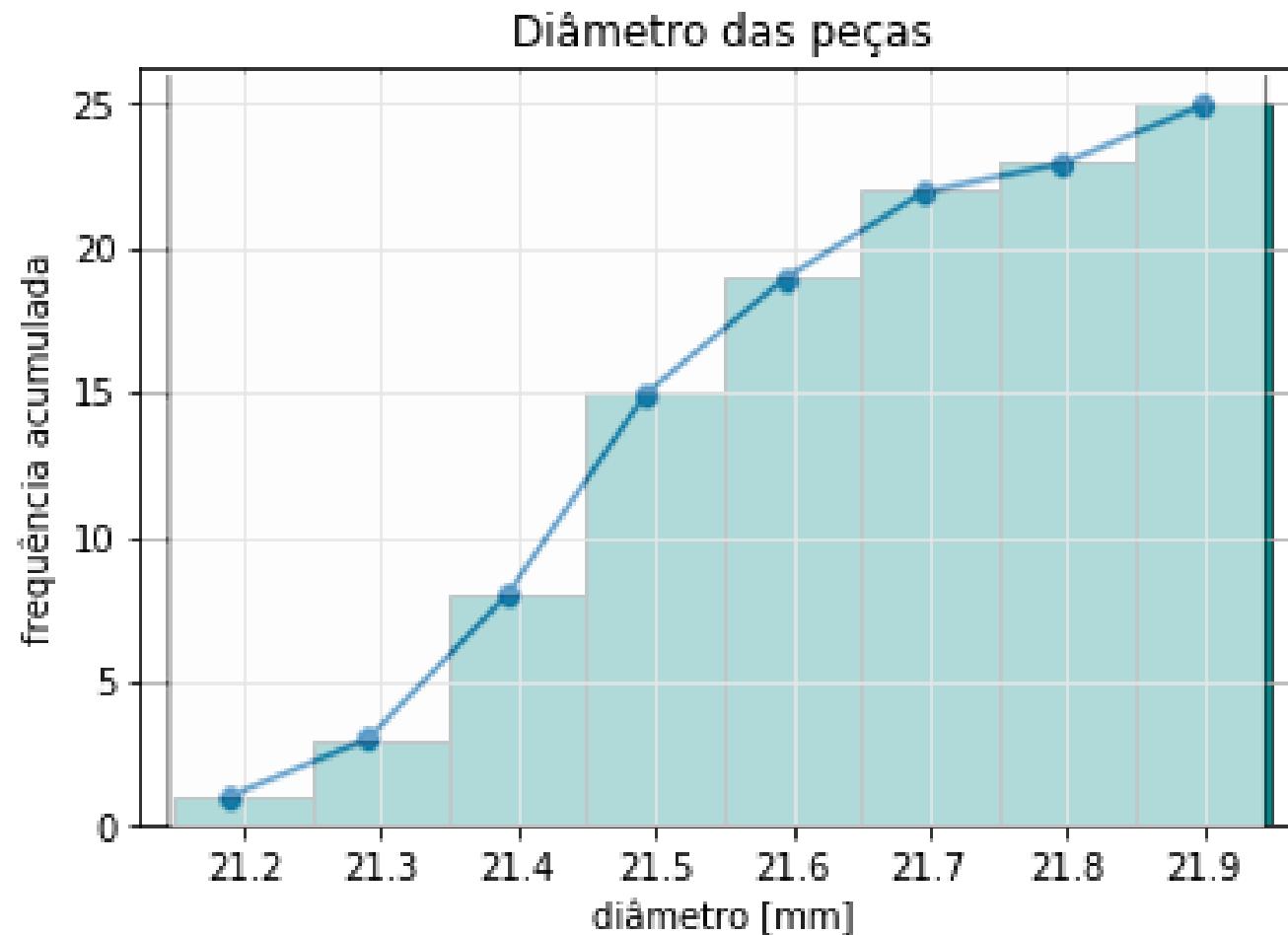


Diâmetro das peças





O histograma também pode ser representado pelo **polígono de frequências**, que é obtido unindo-se os pontos médios das classes de frequência (ou frequências relativas) definidas no histograma. Para completar este polígono de frequências, considera-se as duas classes laterais com frequência nula.



Pode-se também construir o polígono de freqüências acumuladas, somando-se as freqüências acumuladas ao final de cada uma das classes.

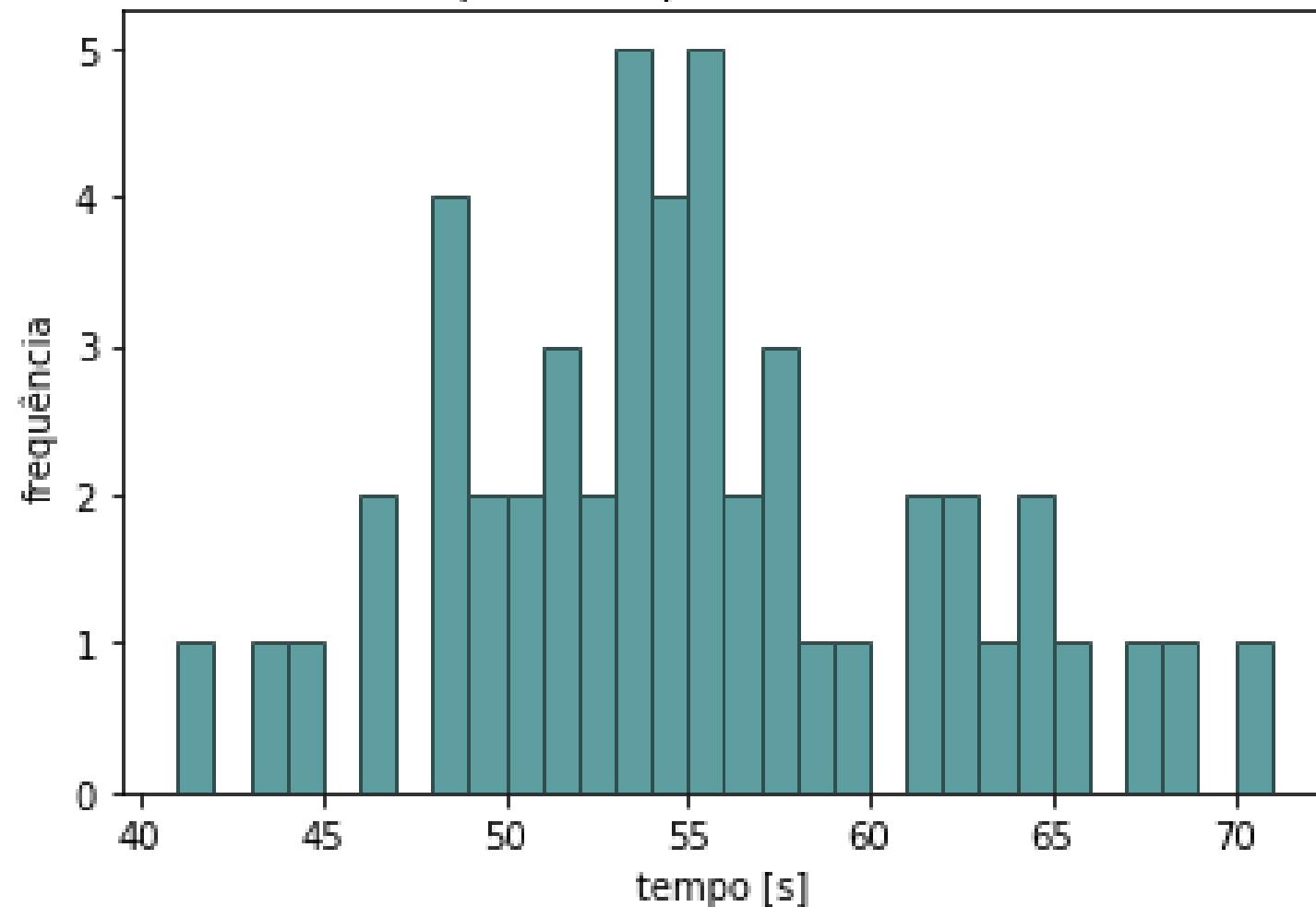
VEJA OUTRO EXEMPLO...

Faça uma descrição do conjunto de 50 observações do tempo necessário para um funcionário preencher um formulário específico.

61	65	43	53	55	51	58	55	59	56
52	53	62	49	68	51	50	67	62	64
53	56	48	50	61	44	64	53	54	55
48	54	57	41	54	71	57	53	46	48
55	46	57	54	48	63	49	55	52	51

Fonte: Costa Neto, P. L. "Estatística", 1^a edição, Editora Edgard Blucher, 1983.

Tempo de resposta de formulário



Histograma de frequência – dados brutos

AGRUPAMENTO DE DADOS

No caso de **variáveis quantitativas contínuas**, algumas vezes é necessário agrupar os dados em **classes de freqüências**, com o intuito de obter uma representação satisfatória.

A freqüência de cada classe é, neste caso, igual a soma das freqüências de todos os valores existentes dentro da classe.

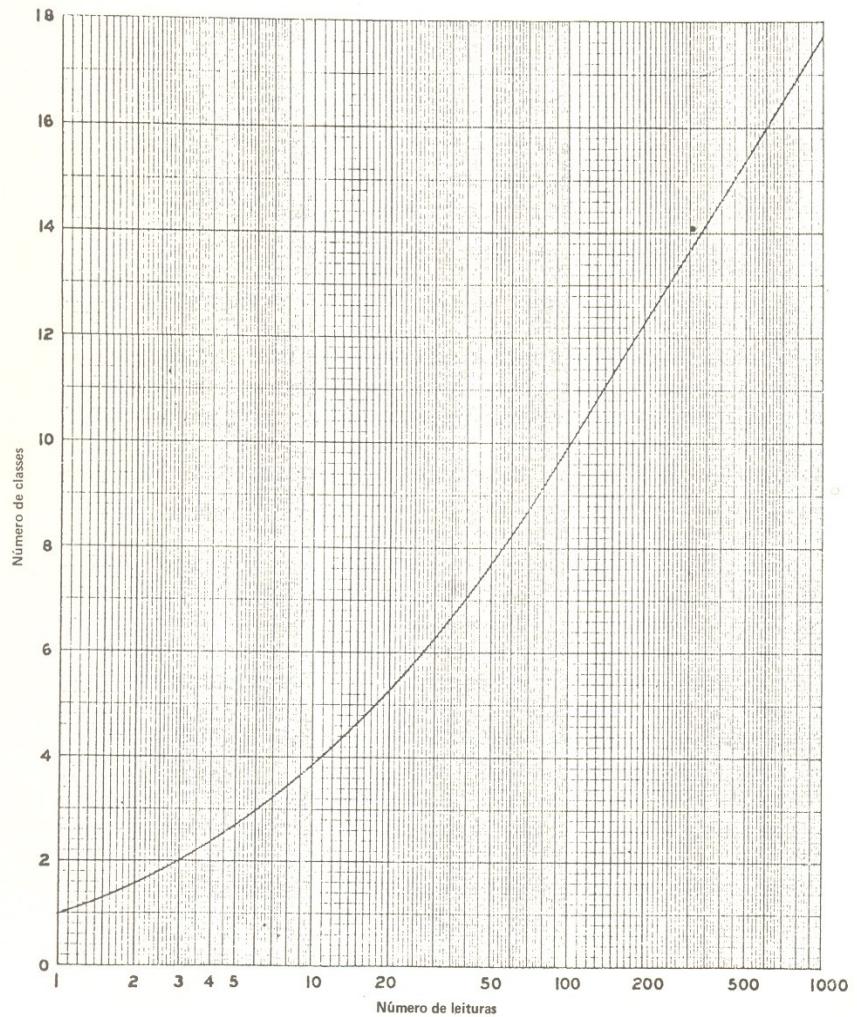
O maior problema neste caso é a definição do número de classes a ser utilizado.

“Se houver muitos intervalos, o resumo não constituirá grande melhoria com relação aos dados brutos. Se houver muito poucos, um grande volume de informação se perderá. Embora não seja necessário, os intervalos são freqüentemente construídos de modo que todos tenham larguras iguais, o que facilita as comparações entre as classes”. (Pagano, 2004, p.11).

O gráfico apresenta uma **sugestão** do número de classes que um conjunto de dados pode ser dividido em função do número de elementos.

O que se faz na prática é tentar variados números de classes e verificar, com a ajuda de um computador, o número ideal para os dados em questão. Além disso, comumente usamos intervalos de classes de iguais amplitudes.

Fonte: Costa Neto, P. L.
“Estatística”, 1^a
edição, Editora Edgard
Blucher, 1983.



CRITÉRIOS PARA A DETERMINAÇÃO DO NÚMERO DE INTERVALOS, k

Raiz quadrada:
(para $25 \leq n \leq 400$) $k = \sqrt{n}$

Log (Sturges):
(para $16 \leq n \leq 572$) $k = -1 + 3,322 \log n$

ln (Milone):
(para $20 \leq n \leq 36$) $k = -1 + 2 \ln n$

Bom senso

Decida a quantidade de classes que
GARANTA observar como os valores se
distribuem.

(n é o número de elementos da amostra).

Fonte: Milone, G. "Estatística Geral e Aplicada". São Paulo: Pioneira Thomson Learning, 2004.

Uma vez definido o número de classes define-se a amplitude h de cada classe. Inicialmente, propõe-se:

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

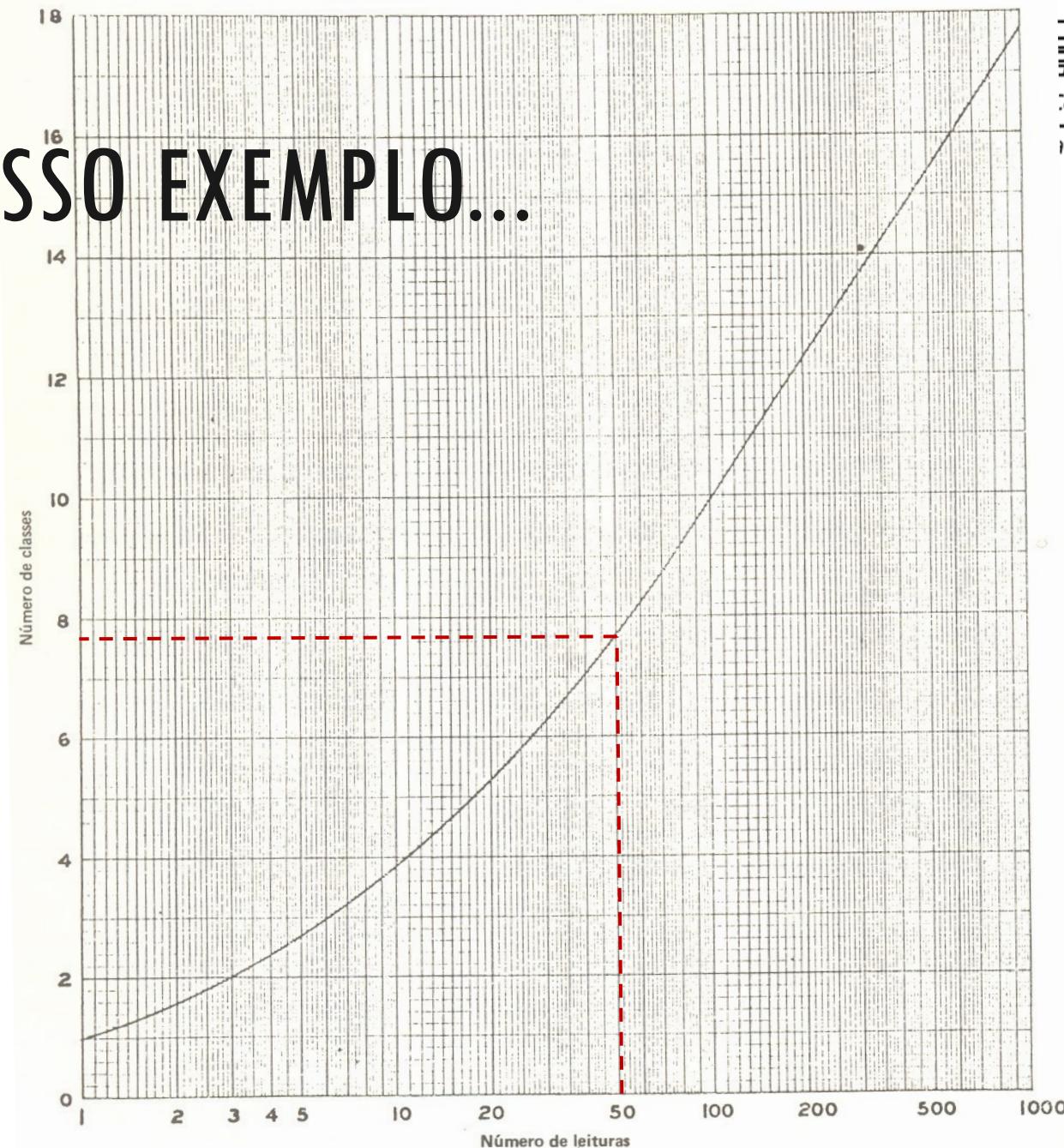
Amplitude R :
diferença entre o maior e o menor dos valores observados

Número de classes k .

VOLTANDO AO NOSSO EXEMPLO...

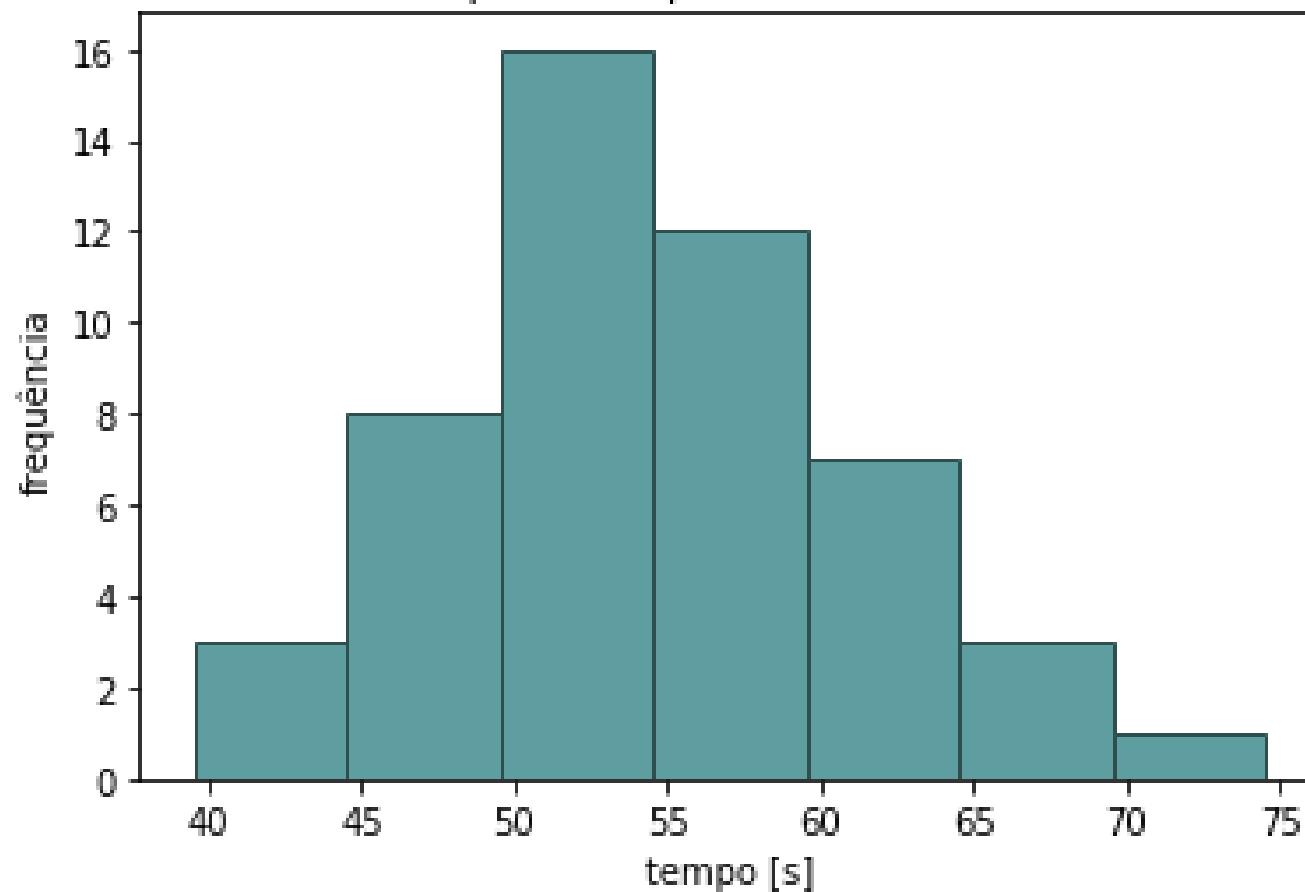
Temos $n=50$,
portanto:

Entre 7-8 classes.
Vamos tentar 7...



CLASSES			f_i
Limites aparentes		Limites reais	
Primeira notação	Segunda notação		
40 45	40-44	39,5 - 44,5	3
45 50	45-49	44,5 - 49,5	8
50 55	50-54	49,5 - 54,5	16
55 60	55-59	54,5 - 59,5	12
60 65	60-64	59,5 - 64,5	7
65 70	65-69	64,5 - 69,5	3
70 75	70-74	69,5 - 74,5	1
			50

Tempo de resposta de formulário



Histograma de frequências – dados agrupados

CARACTERÍSTICAS NUMÉRICAS DE UMA DISTRIBUIÇÃO DE FREQUÊNCIAS

Além da descrição gráfica muitas vezes é necessário descrever algumas características da distribuição de frequências através das ditas *medidas descritivas*.

Estas medidas, **se calculadas a partir de dados populacionais**, são denominadas **parâmetros** e **se calculadas a partir de dados amostrais** são denominadas **estimadores ou estatísticas**.

Usualmente empregam-se as medidas de posição (*tendência central*), dispersão, e assimetria, sendo que as duas primeiras são as mais empregadas.

MEDIDAS DE POSIÇÃO

As medidas de posição são empregadas para localizar a distribuição de freqüências sobre o eixo de variação da variável em estudo.

Usualmente emprega-se para este fim a **média aritmética**, **mediana**, **moda**. Por critérios diferentes, todas indicam o centro da distribuição de freqüências, sendo por isso também denominadas de medidas de **tendência central**.

MÉDIA ARITMÉTICA

\bar{x} (AMOSTRA) OU μ (POPULAÇÃO)

Dado um conjunto de dados x_1, x_2, \dots, x_n

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

Caso os dados sejam apresentados na forma de uma tabela de frequências a média pode ser calculada pela relação:

$$\bar{x} = \frac{\sum_{i=1}^k x_i f_i}{n} = \sum_{i=1}^k x_i p_i$$

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

A média aritmética é a soma de todos os valores observados da variável dividida pelo número total de observações.

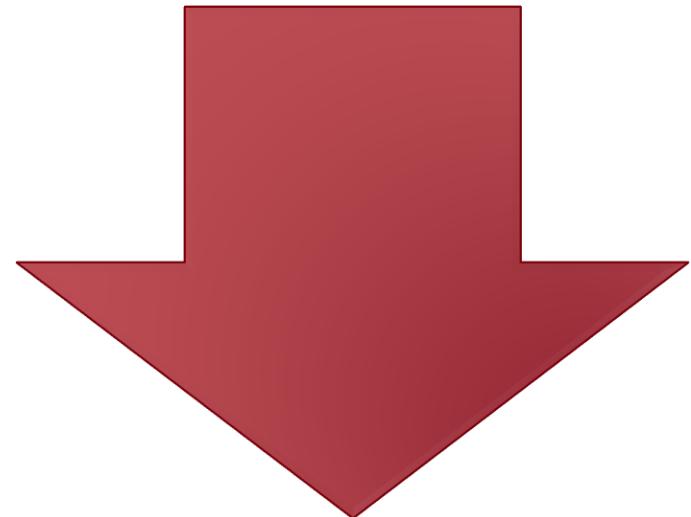
Considere os valores correspondentes à ajuda de custos de estagiários paga por empresas da área de engenharia.

Empresa	Salário dos Estagiários (R\$)												
A	200	200	200	840	200	200	300	200	300	350	700	350	950
B	200	200	840	350	300	300	200	200	950	200			

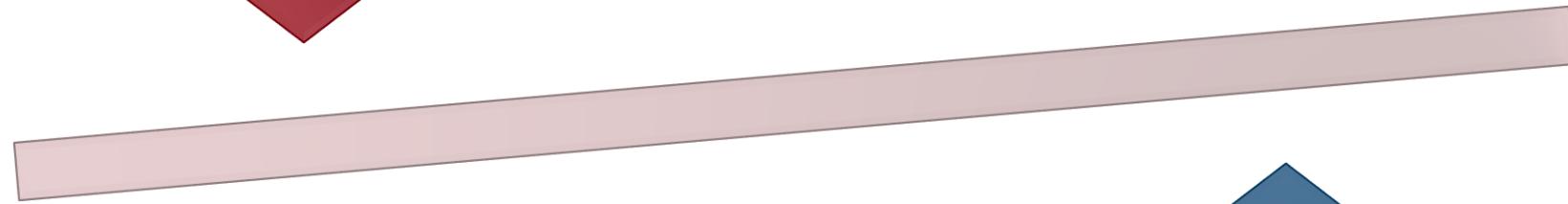
$$\bar{x}_A = \frac{R\$4990,00}{13} = R\$383,85$$

$$\bar{x}_B = \frac{R\$3740,00}{10} = R\$374,00$$

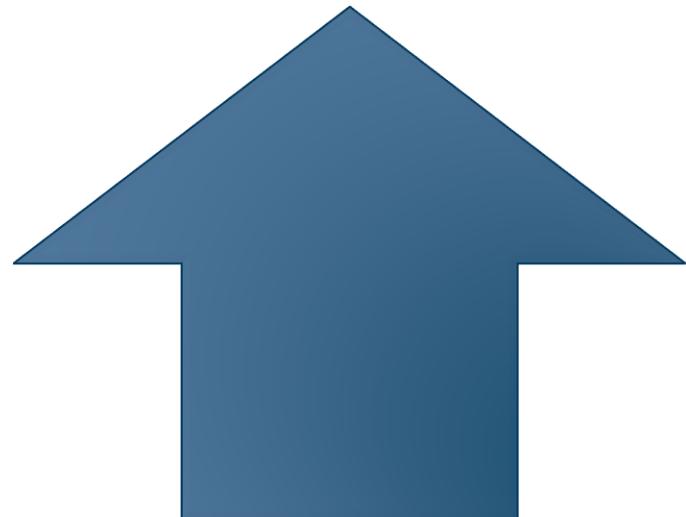
O salário médio dos 13 estagiários da empresa A é R\$ 383,85 e dos 10 estagiários da empresa B é R\$ 374,00.



A média caracteriza o centro da distribuição de frequências, sendo, por isso, uma medida de posição.



Em uma analogia de massas, a média corresponde ao centro de gravidade da distribuição de frequências.



SOBRE A MÉDIA ARITMÉTICA

Depende de todas as observações;
Qualquer modificação nos dados fará com que a média fique alterada – afetada por valores extremos (*outliers*);

É única em um conjunto de dados e nem sempre tem existência real, ou seja, nem sempre é igual a um determinado valor observado;

A soma da diferença de cada valor observado em relação à média é zero

$$\sum_i (x_i - \bar{x}) = 0$$



MÉDIA APARADA OU TRUNCADA

Trimmed mean

Calcula-se a média depois de excluir um número fixo m de valores de cada ponta do conjunto de dados.

$$\bar{x} = \frac{x_{m+1} + x_2 + \dots + x_{n-m}}{n - 2m} = \frac{\sum_{i=m+1}^{n-m} x_i}{n - 2m}$$

Por exemplo, em uma competição internacional de mergulho, as notas máxima e mínima dos 5 juízes são descartadas, e a nota final é dos 3 restantes.

MÉDIA ARITMÉTICA PONDERADA

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

A **média aritmética ponderada** é utilizada quando cada valor do conjunto **possui um peso diferente**.

E AGORA?

Em uma empresa, o salário médio dos estagiários é de R\$500,00. Os salários médios pagos aos estagiários nível I e II são R\$520,00 e R\$420,00, respectivamente. Pode-se dizer, então, que:

- A. O número de estagiários nível I é o dobro do número de estagiários nível II .
- B. O número de estagiários nível I é o triplo do número de estagiários nível II.
- C. O número de estagiários nível I é o quádruplo do número de estagiários nível II.
- D. O número de estagiários nível II é o triplo do número de estagiários nível I.
- E. O número de número de estagiários nível II é o quádruplo do número de estagiários nível I.

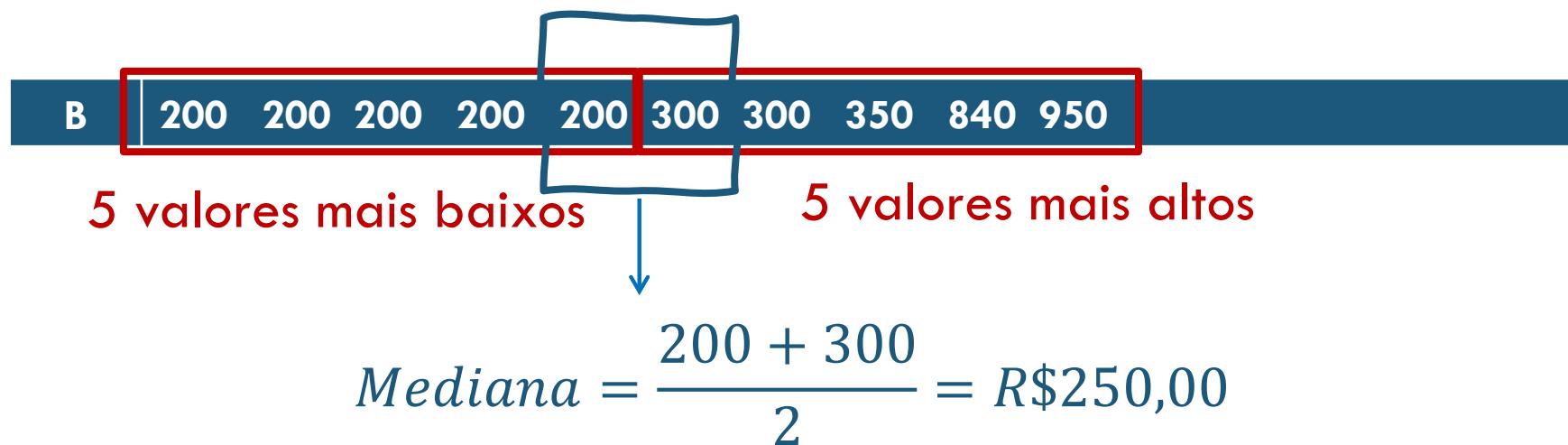
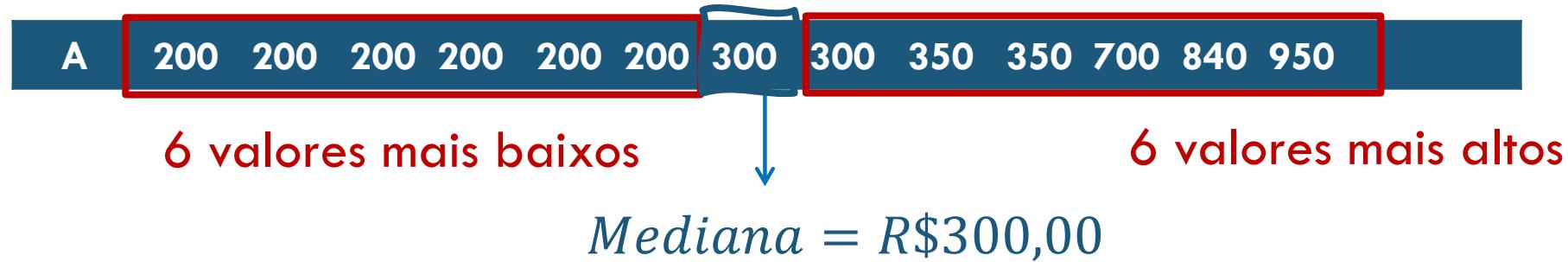
$$\bar{x} = \frac{\bar{x}_I n_I + \bar{x}_{II} n_{II}}{n_I + n_{II}}$$

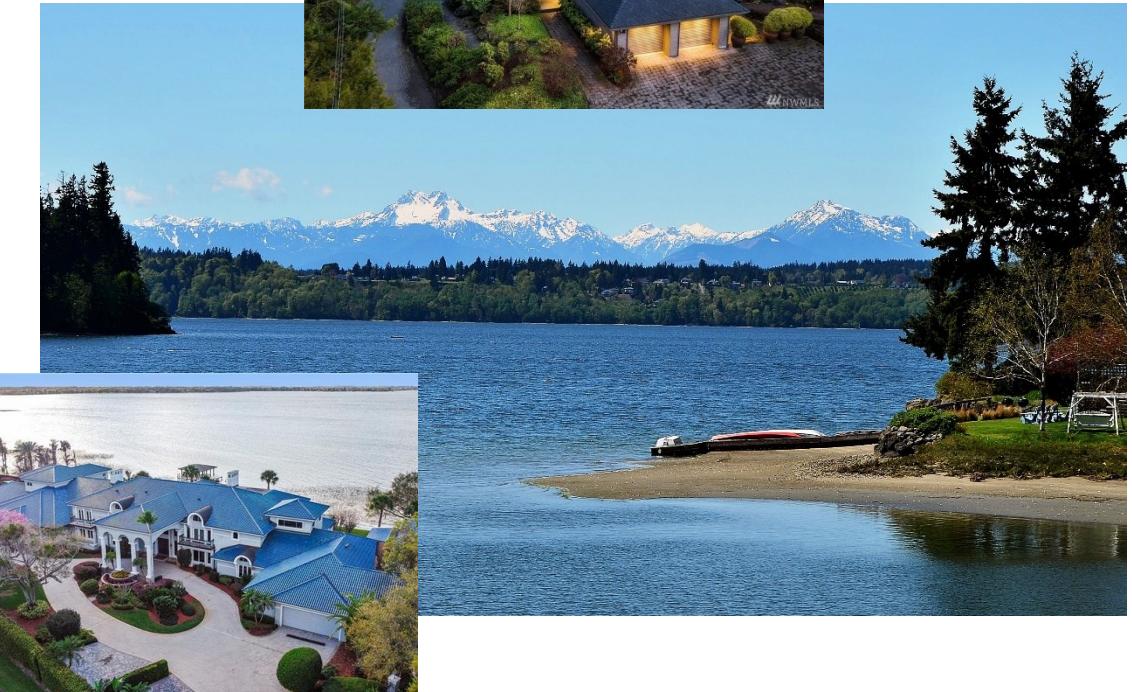
MEDIANA

A mediana é o número central em uma lista de dados classificada. É o valor que divide o conjunto de dados em dois subconjuntos: 50% do valores estão abaixo da mediana e 50% dos valores estão acima da mediana.



Considere o exemplo anterior. Os valores correspondentes à ajuda de custos de estagiários paga por empresas da área de engenharia foram colocados em ordem crescente,





Windermere

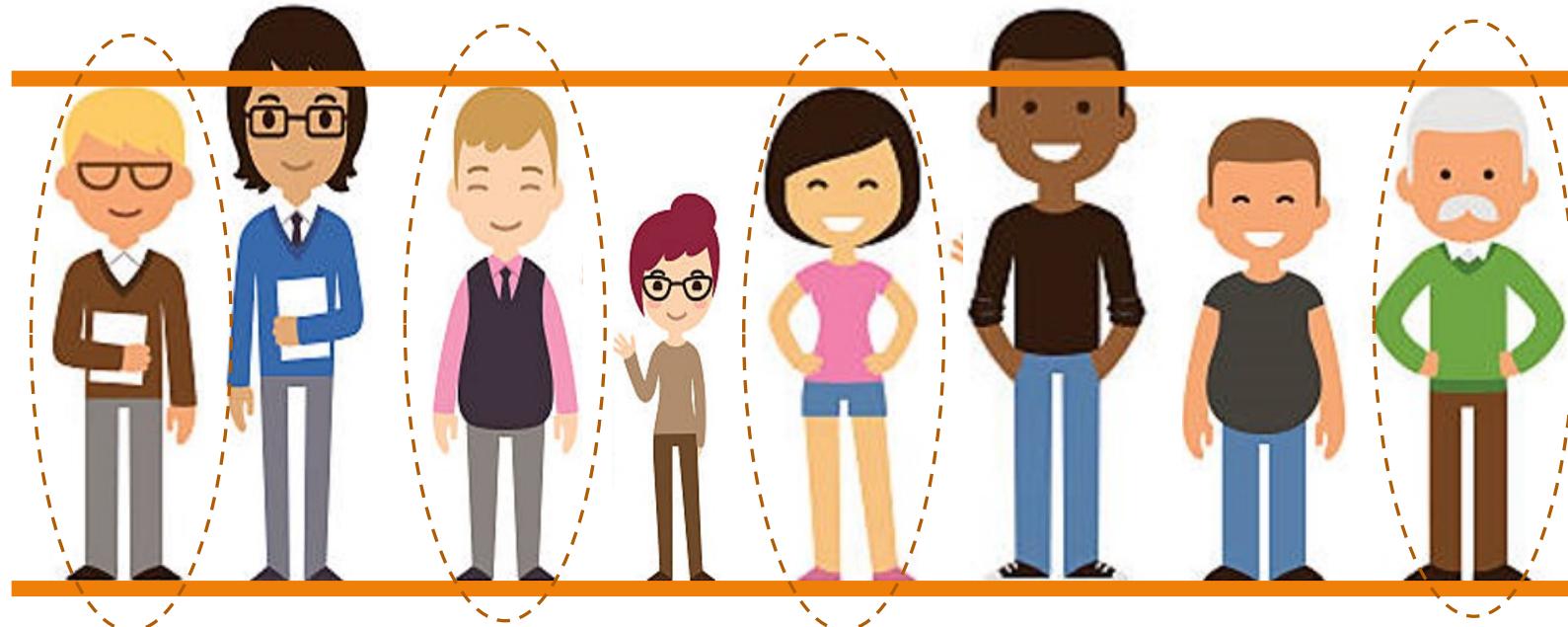


Medina

Se fizesse a média e a mediana dos rendimentos familiares nestes dois lugares, você acha que obteria aproximadamente o mesmo valor?

MODA

Moda de um conjunto de valores é o valor de máxima frequência. A distribuição será unimodal no caso de apresentar uma só moda e multimodal se apresentar várias modas.



Considerando ainda o exemplo anterior. As frequências de cada valor correspondente à ajuda de custos de estagiários paga por empresas da área de engenharia são mostradas abaixo,

Empresa A

Salário	200	300	350	700	840	950
Frequência	6	2	2	1	1	1

Empresa B

Salário	200	300	350	840	950
Frequência	5	2	1	1	1

A moda das empresas A e B é R\$200,00.

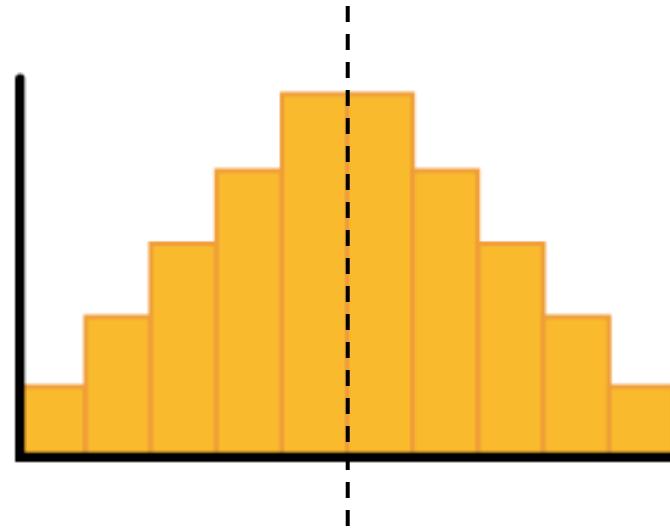
Medida de tendência central	Definição	Empresa A (R\$)	Empresa B (R\$)
Média	Soma de dados/Número total de dados	383,85	374,00
Mediana	Valor médio da lista ordenada	300,00	250,00
Moda	Valor mais frequente	200,00	200,00

Média > Mediana

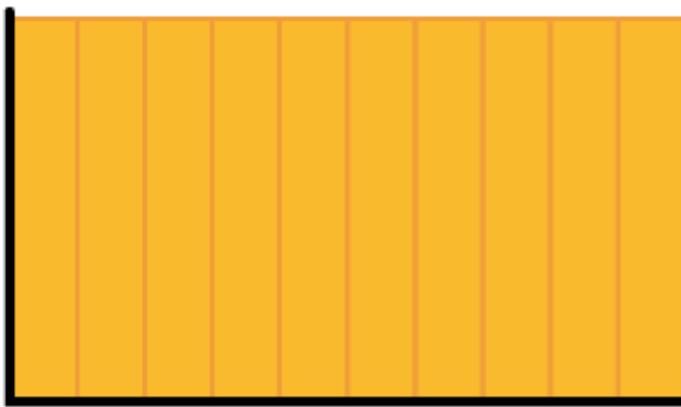
A média é fortemente afetada por poucos altos salários em cada conjunto de dados. A média é sensível a valores extremos. Por esse motivo, a mediana é chamada de estimativa robusta de localização.

Media, mediana e moda geralmente proporcionam informações diferentes. Não existe uma regra clara e simples de qual medida se deve usar em uma dada situação.

Média ≡ Mediana ≡ Moda

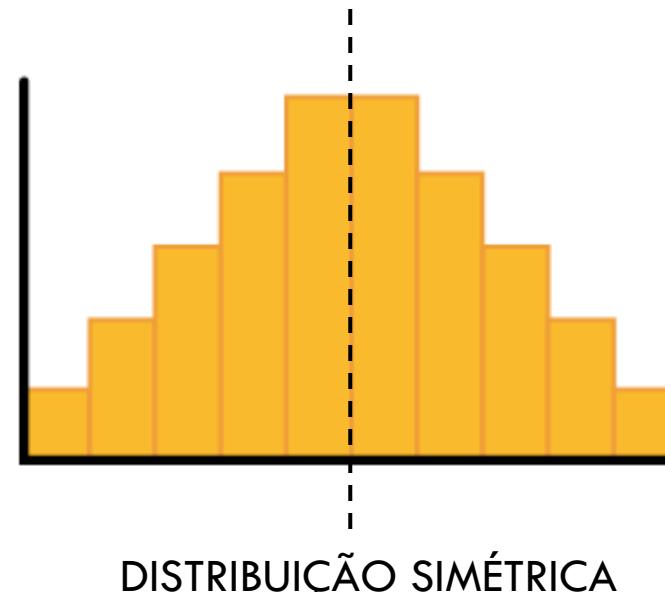


DISTRIBUIÇÃO SIMÉTRICA



DISTRIBUIÇÃO UNIFORME

Média \equiv Mediana \equiv Moda

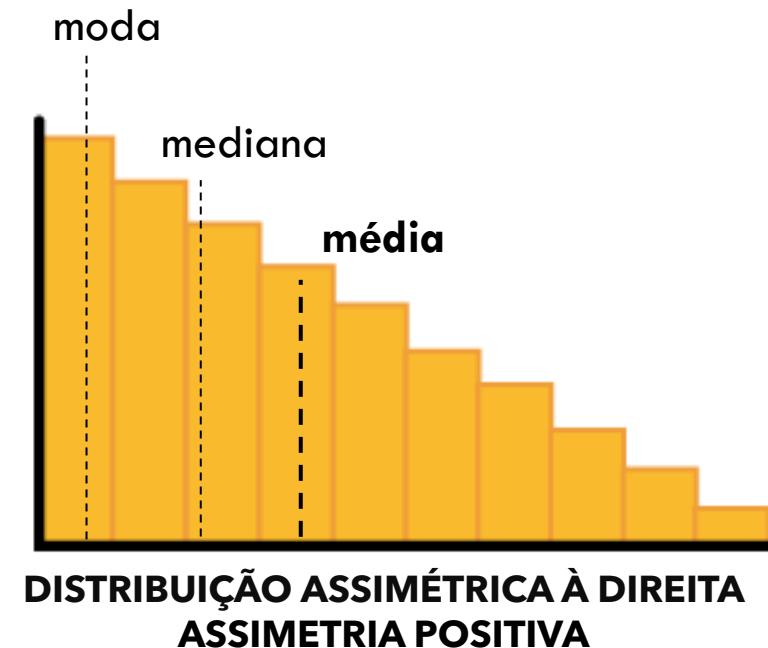


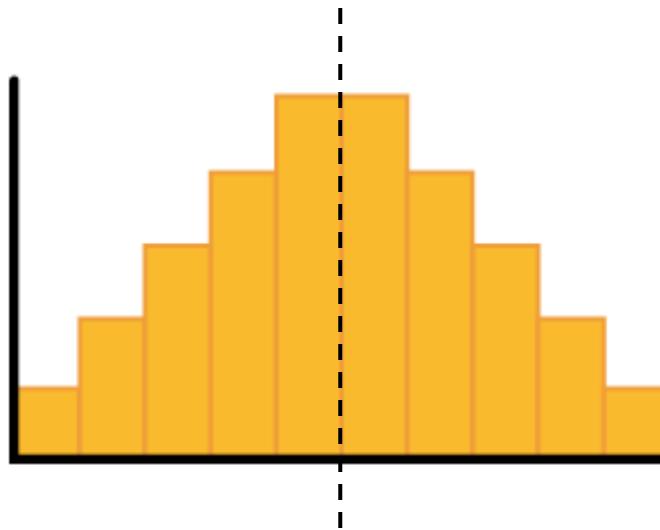
Em uma distribuição assimétrica à direita, a cauda do lado direito é mais longa, e a maior parte da distribuição está concentrada à esquerda.

Os valores extremos no extremo superior puxam a média para a direita.
Isso faz com que a média seja maior que a mediana.

Além disso, em uma distribuição com assimetria positiva, a moda é tipicamente menor que a mediana, já que é influenciada pelo agrupamento de valores em direção ao extremo inferior da escala.

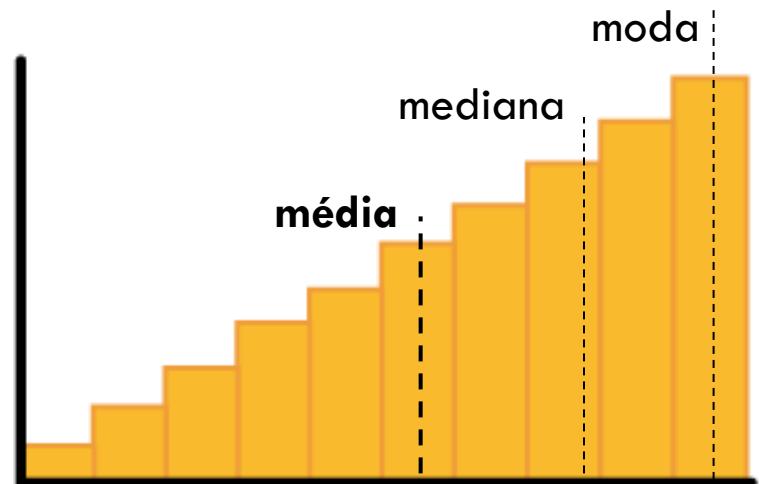
Pode-se medir a assimetria de uma distribuição de dados (Skewness).





DISTRIBUIÇÃO SIMÉTRICA

Pode-se medir a assimetria de uma distribuição de dados (Skewness).



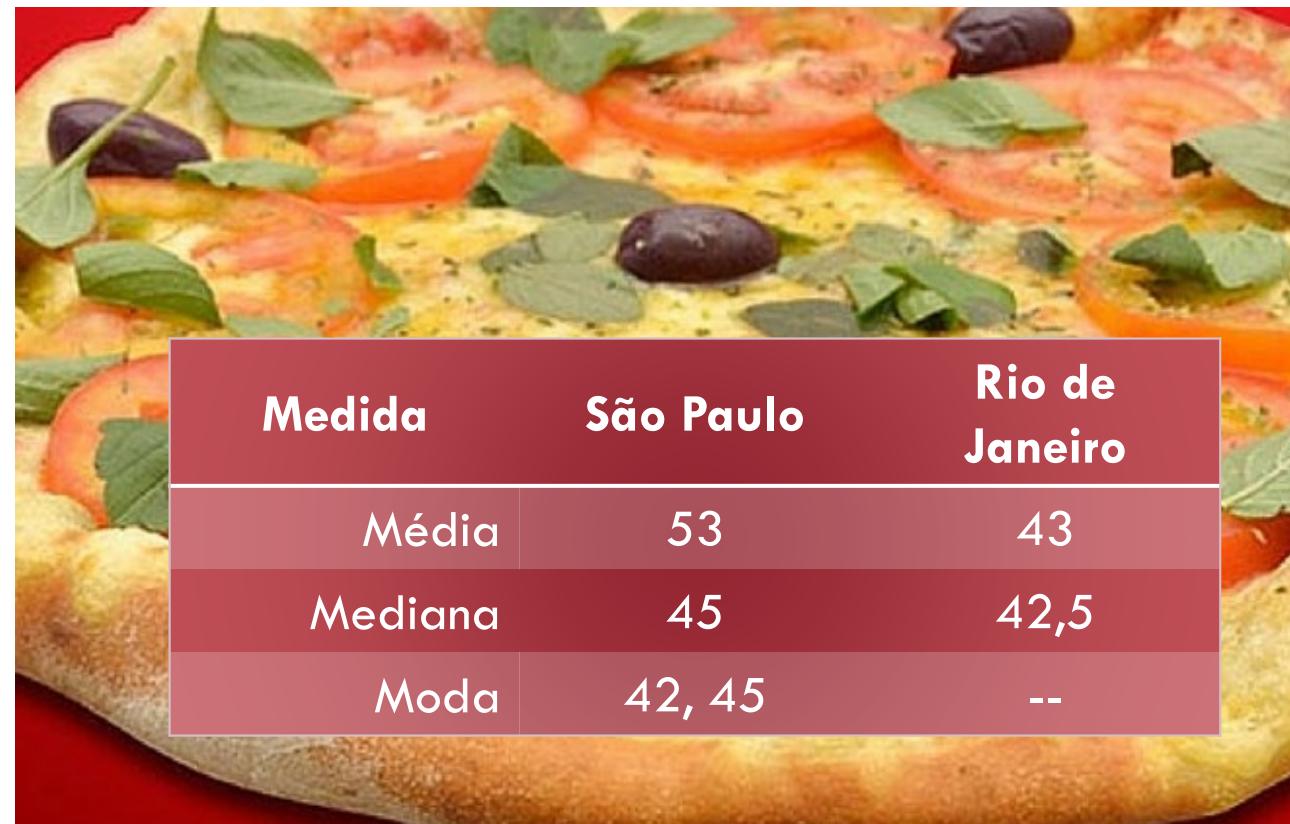
DISTRIBUIÇÃO ASSIMÉTRICA À ESQUERDA
ASSIMETRIA NEGATIVA

Em uma distribuição assimétrica à esquerda, a cauda do lado esquerdo é mais longa, e a maior parte da distribuição está concentrada à direita.

Os valores extremos no extremo inferior puxam a média para a esquerda. Isso faz com que a média seja menor que a mediana. Além disso, em uma distribuição com assimetria negativa, a moda é tipicamente maior que a mediana, já que é influenciada pelo agrupamento de valores em direção ao extremo superior da escala.

FAÇA PARA ESTUDAR: PREÇO DA PIZZA MARGHERITA

Índice	São Paulo	Rio de Janeiro
1	39	38
2	40	39
3	41	40
4	42	41
5	42	42
6	45	43
7	45	44
8	47	46
9	49	48
10	50	49
11	143	



Não há moda quando todos os valores observados aparecem o mesmo número de vezes em um conjunto de dados.

VAMOS TREINAR

Que medida de tendência central é mais adequada para se analisar cada um dos casos abaixo? Justifique.

- Um estudante faz quatro provas em uma disciplina de cálculo. Suas notas são 8,8; 7,5; 9,5 e 10,0.
- Uma grande construtora publica os preços de venda de seus apartamentos disponíveis na grande São Paulo.
- Em uma maratona existiam duas categorias de corredores, masculinos e femininos. A tabela seguinte mostra a distribuição de frequências para os dados.

Sexo	Frequência
Masculino	4239
Feminino	964

MEDIANA VS MÉDIA

A mediana vence da média quando...

A mediana é uma escolha melhor quando o indicador pode ser afetado por alguns valores discrepantes (outliers). A mediana é o valor médio que não é afetado por esses valores extremos. Isso nos dá uma estimativa melhor dos valores típicos. No entanto, isso significa que a presença de valores relativamente mais altos está oculto.

Média vence da mediana quando...

A média é uma escolha melhor quando não há valores extremos que possam afetá-la. É um resumo melhor, porque as informações de todas as observações são incluídas na média e não na mediana, que é apenas o valor do meio.

E A MODA?

A moda faz sentido, por exemplo, quando não temos um conjunto de dados com valores numéricos, o que é necessário no caso da média e da mediana.

Porém, ela não é útil quando nossa distribuição é uniforme; ou seja, as frequências de todos os grupos são semelhantes. Nesses casos, a moda não fornece nenhuma informação útil.

Às vezes, a moda pode aparecer no final da distribuição, o que não é necessariamente no centro ou perto do centro de uma distribuição.

E AGORA?

Considere a idade de cada membro do grupo de avós e netos que brincam em uma praça,

1 1 1 2 2 2 3 3 3 3 3 50 52 55 55 57 58 59 59 60 61 65

Qual medida melhor representa a média de idade dos frequentadores do parque?

Média

$$\frac{655}{22} = 29,77$$

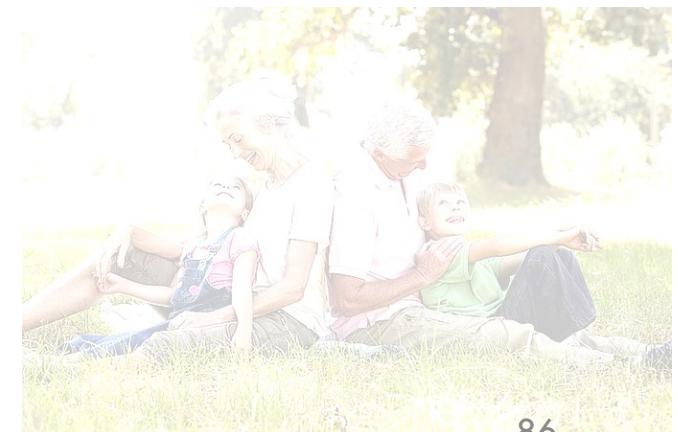
Mediana

$$\frac{(50 + 3)}{2} = 26,5$$

ou

Moda?

3



REGRESSÃO À MÉDIA

Mário e Maria estão entrevistando candidatos a um emprego de vendedor de sapatos. Acaba de sair um candidato de 2 metros de altura. Maria diz: “aposto que o próximo candidato é mais baixo”. Mário retruca: “a altura do candidato nada tem a ver com o histórico dos candidatos anteriores”. Quem tem razão?

Mlodinow, no livro *O andar do bêbado*, conta a história de Daniel Kahneman, psicólogo que em 2002 ganhou o Nobel de Economia: “Como não se escolhe trabalho no começo da carreira, nos anos 1960 ele foi ensinar aos instrutores da Aeronáutica israelense que recompensar funciona melhor do que punir erros. Foi interrompido por um dos instrutores que o ouvia. Ele dizia que muitas vezes elogiou a manobra de um aluno e, na vez seguinte, o sujeito se saiu muito pior. E que, quando gritou com a besta que havia quase acabado de destruir o avião, ela melhorava em seguida. Os outros instrutores concordaram.” Estaria o psicólogo errado?

PARA VOCÊ PENSAR...

- Se o salário médio de todos os empregados homens da companhia A excede o salário médio de todos os empregados homens da companhia B, e se o salário médio de todas as empregadas mulheres da companhia A excede o salário médio de todas as empregadas mulheres da companhia B, decorre daí que o salário médio da totalidade dos empregados da companhia A excede o salário médio da totalidade dos empregados da companhia B? Explique sua resposta.
- Durante as três semanas antes do Natal, 12 pessoas fizeram compras, em média, em 5,75 lojas de roupas. É possível que ao menos sete delas tenham feito compras em pelo menos 10 lojas?

MEDIDAS DE DISPERSÃO

A informação fornecida pelas medidas de posição necessita, em geral, ser complementada pelas medidas de dispersão.

"Fenômenos que envolvem análises estatísticas caracterizam-se por suas semelhanças e variabilidades."

As medidas de dispersão são utilizadas para indicar o quanto os dados se apresentam dispersos em torno da região central, caracterizando, portanto, o grau de variação existente no conjunto de valores.

As medidas de dispersão usualmente utilizadas são **amplitude, variância, desvio padrão, distância interquartil**.

AMPLITUDE

A amplitude, já mencionada anteriormente, é definida como a diferença entre o maior e o menor valores do conjunto de dados:

$$R = X_{max} - X_{min}$$

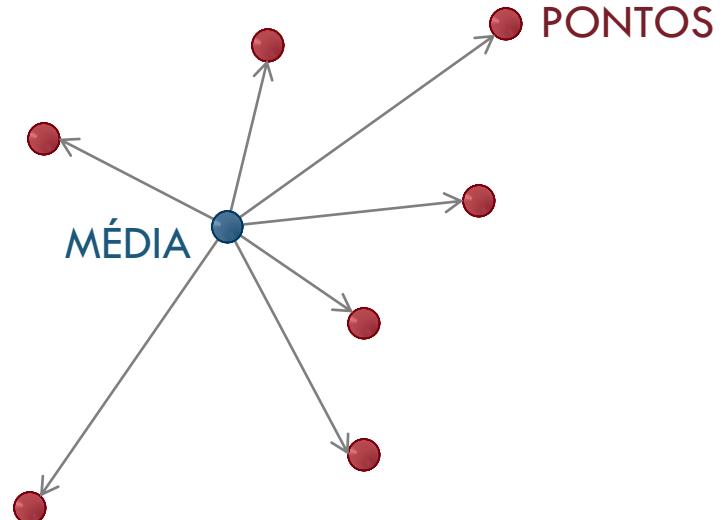
Embora o valor de R esteja relacionado com a dispersão dos dados, por usar apenas dois valores, a amplitude contém relativamente pouca informação quanto à dispersão.



DESVIO PADRÃO

As estimativas de variação mais utilizadas são baseadas nas diferenças, ou desvios, entre a estimativa de localização e o dado observado.

Desvio padrão mede a variação em um conjunto de dados através da determinação do quanto, em média, cada valor está distante da média do conjunto.



Vamos ilustrar o
desvio com relação
à média.

O robô foi programado para deslocar o braço 5 vezes na direção x , do ponto A ao ponto B. Os deslocamentos medidos foram:

184, 186, 193, 193 e 199 mm.

Determine os desvios em relação à média.



$$\bar{x} = \frac{\sum x}{n} = \frac{184 + 186 + 193 + 193 + 199}{5} = 191 \text{ mm}$$

Deslocamento	Desvio em relação à média	Quadrado do desvio
184	-7	49
186	-5	25
193	2	4
193	2	4
199	8	64
\sum	0	146

VARIÂNCIA

A variância de um conjunto de dados é a média dos quadrados das diferenças dos valores em relação à sua média, isto é:

$$\sigma_x^2 = \sum_{i=1}^n \frac{(x_i - \mu)^2}{n}$$

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

Nesta formulação os valores mais distantes da média exercem maior influência sobre o valor da variância.

Como exemplo, calcula-se a variância do deslocamento do robô,

$$\bar{x} = \frac{\sum x}{n} = \frac{184 + 186 + 193 + 193 + 199}{5} = 191 \text{ mm}$$

$$s_x^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{146}{4} = 36,5 \text{ mm}^2$$



Enquanto as unidades da média, moda, mediana e desvio padrão possuem interpretação (por exemplo: m, m², m³, s, kg, filhos, anos), a variância não admite interpretação prática por causa das unidades (por exemplo: m², m⁴, m⁶, s², kg², filhos², anos², respectivamente).

DESVIO PADRÃO

É definido como a raiz quadrada positiva da variância, representado pelo símbolo s_x :

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

O desvio padrão se expressa na mesma unidade da variável, sendo, por isso, de maior interesse que a variância nas aplicações práticas.

No exemplo: $s_x = \sqrt{36,5} = 6,04 \text{ mm}$



QUARTIL

São valores dados a partir do conjunto de observações ordenado de forma crescente, que dividem a distribuição em quatro partes iguais.

O terceiro quartil, Q3: valor que deixa 75% das observações abaixo e 25% acima.

O segundo quartil, Q2: é a mediana, deixa 50% das observações abaixo e 50% das observações acima;

O primeiro quartil, Q1: valor que deixa 25% das observações abaixo e 75% acima;



ACHE OS QUARTIS DOS SEGUINTEIS CONJUNTOS DE DADOS

$$A = \{3, 7, 8, 5, 12, 14, 21, 15, 18, 14\}$$

$$B = \{3, 7, 8, 5, 12, 14, 21, 13, 18\}$$

$$C = \{19, 26, 25, 37, 32, 28, 22, 23, 29, 34, 39, 31\}$$

Vamos fazer para os dados A , B e você faz o C quando estiver estudando...

PASSOS DO ALGORITMO PARA ENCONTRAR QUARTIS

1. Classifique o conjunto de dados em ordem crescente e verifique o número total de entradas no conjunto de dados n .

2. Se o número de entradas for par:

- 1.** Calcule a mediana (Q_2) tirando a média dos dois valores do meio.
- 2.** Divida o conjunto de dados em duas metades: a primeira metade contendo as menores $n/2$ entradas e a segunda metade contendo as maiores $n/2$ entradas.
- 3.** Calcule Q_1 como a mediana da primeira metade.
- 4.** Calcule Q_3 como a mediana da segunda metade.

Os valores calculados de Q_1 , Q_2 e Q_3 representam o primeiro quartil, mediana (segundo quartil) e terceiro quartil, respectivamente.

QUARTIS

3, 7, 8, 5, 12, 14, 21, 15, 18, 14

Organize os dados em ordem crescente (a mediana é a referência do quartil)



Ache a mediana

$n = 10$

Q2 ou 2º Quartil: O segundo quartil corresponde a 50% da distribuição

$$Q_{50\%} = Q_2 = \frac{12+14}{2} = 13$$

Portanto, a metade inferior dos dados é: [3, 5, 7, 8, 12] (Observe que o valor 12 usado para calcular a mediana está incluído) e o **1º Quartil ou Q1** é a sua mediana, que corresponde a 25% da distribuição

$$Q_{25\%} = Q_1 = 7$$

Da mesma forma, a metade superior dos dados é: [14, 14, 15, 18, 21], então **3º Quartil ou Q3**, que corresponde a 75% da distribuição é dado por:

$$Q_{75\%} = Q_3 = 15$$

NUMPY...

```
A = [3, 7, 8, 5, 12, 14, 15, 21, 14, 18]  
  
p25=np.quantile(y, .25, interpolation='midpoint')  
p50=np.quantile(y, 0.5, interpolation='midpoint')  
p75=np.quantile(y, 0.75, interpolation='midpoint')  
print('Percentis 25,50,75: {:.4.2f}, {:.4.2f} e {:.4.2f}'.format(p25, p50, p75))
```

Percentis 25,50,75: 7.50, 13.00 e 14.50

O Python inclui os valores 12,14 usados para calcular a mediana (Q2) nos cálculos do primeiro e terceiro quartis.

1º Quartil ou Q1: é a mediana do conjunto [3,5,7,8,12,14] , que corresponde a 25% da distribuição
$$Q_{25\%} = Q_1 = 7,5$$

3º Quartil ou Q3: é calculado com a parte superior dos dados é: [12,14,14,15,18,21] , então, que corresponde a 75% da distribuição é dado por:

$$Q_{75\%} = Q_3 = 14,5$$

PASSOS DO ALGORITMO PARA ENCONTRAR QUARTIS

1. Classifique o conjunto de dados em ordem crescente e verifique o número total de entradas no conjunto de dados n .
2. Se o número de entradas for par: (...)
- 3. Se o número de entradas for ímpar:**
 1. Calcule a mediana (Q_2) como o valor do meio.
 2. Divida o conjunto de dados em duas metades: a primeira metade contendo as menores $(n - 1)/2$ entradas e a segunda metade contendo as maiores $(n - 1)/2$ entradas.
 3. Calcule Q_1 como a mediana da primeira metade.
 4. Calcule Q_3 como a mediana da segunda metade.

QUARTIS

3, 7, 8, 5, 12, 14, 21, 13, 18

Organize os dados em ordem crescente (a mediana é a referência do quartil)



$n = 9$

Ache a mediana

Q2 ou 2º Quartil: O segundo quartil corresponde a 50% da distribuição

$$Q_{50\%} = Q_2 = 12$$

Portanto, a metade inferior dos dados é: [3, 5, 7, 8] e o **1º Quartil ou Q1** é a sua mediana, que corresponde a 25% da distribuição

$$Q_{25\%} = Q_1 = 6$$

Da mesma forma, a metade superior dos dados é: [13, 14, 18, 21], então **3º Quartil ou Q3**, que corresponde a 75% da distribuição é dado por:

$$Q_{75\%} = Q_3 = 16$$

NUMPY...

```
A = [3, 7, 8, 5, 12, 14, 21, 13, 18]

p25=np.quantile(y, .25, interpolation='midpoint')
p50=np.quantile(y, 0.5, interpolation='midpoint')
p75=np.quantile(y, 0.75, interpolation='midpoint')
print('Percentis 25,50,75: {:.4.2f}, {:.4.2f} e {:.4.2f}'.format(p25, p50, p75))
```

Percentis 25,50,75: 7.00, 12.00 e 14.00

O Python inclui o valor 12 usado para calcular a mediana (Q2) nos cálculos do primeiro e terceiro quartis.

1º Quartil ou Q1: é a mediana do conjunto [3,5,7,8,12] , que corresponde a 25% da distribuição

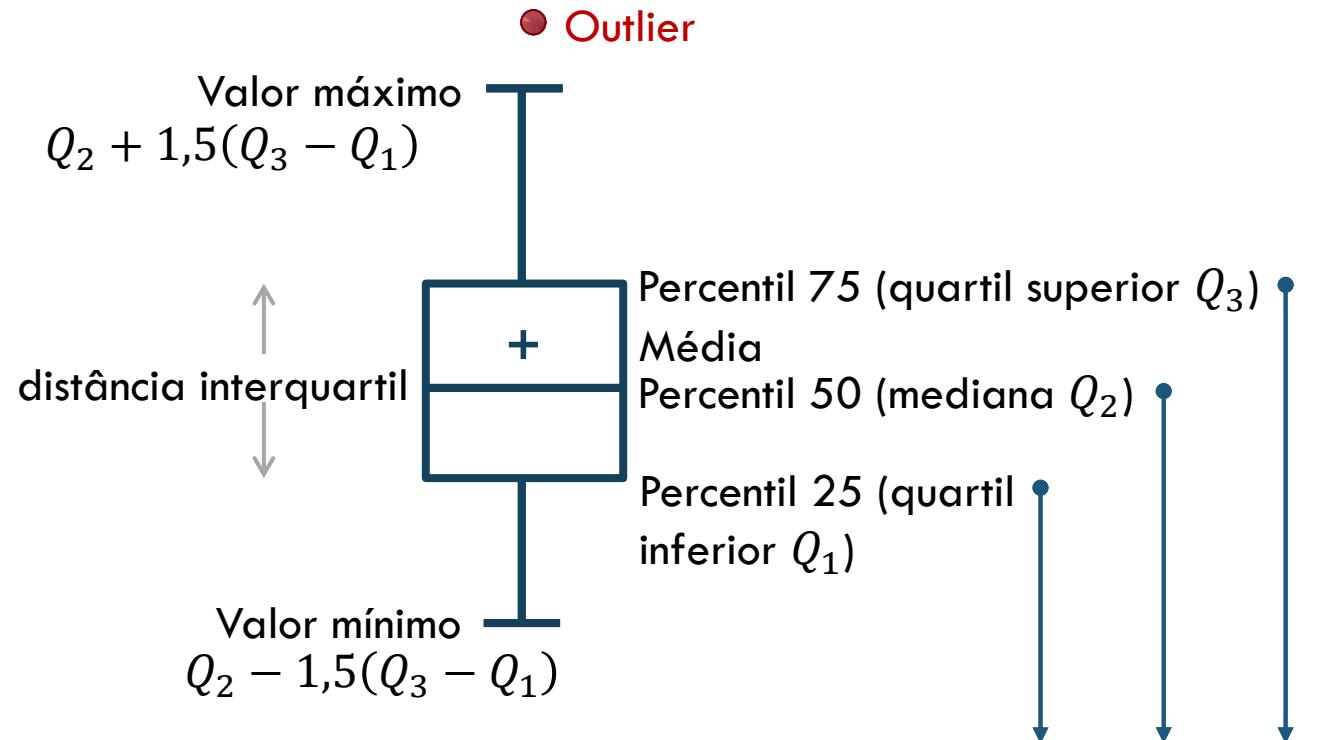
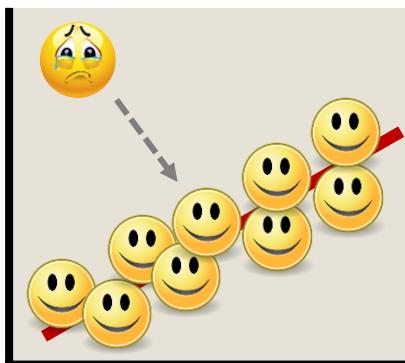
$$Q_{25\%} = Q_1 = 7$$

3º Quartil ou Q3: é calculado com a parte superior dos dados é: [12,13,14,18,21] , então, que corresponde a 75% da distribuição é dado por:

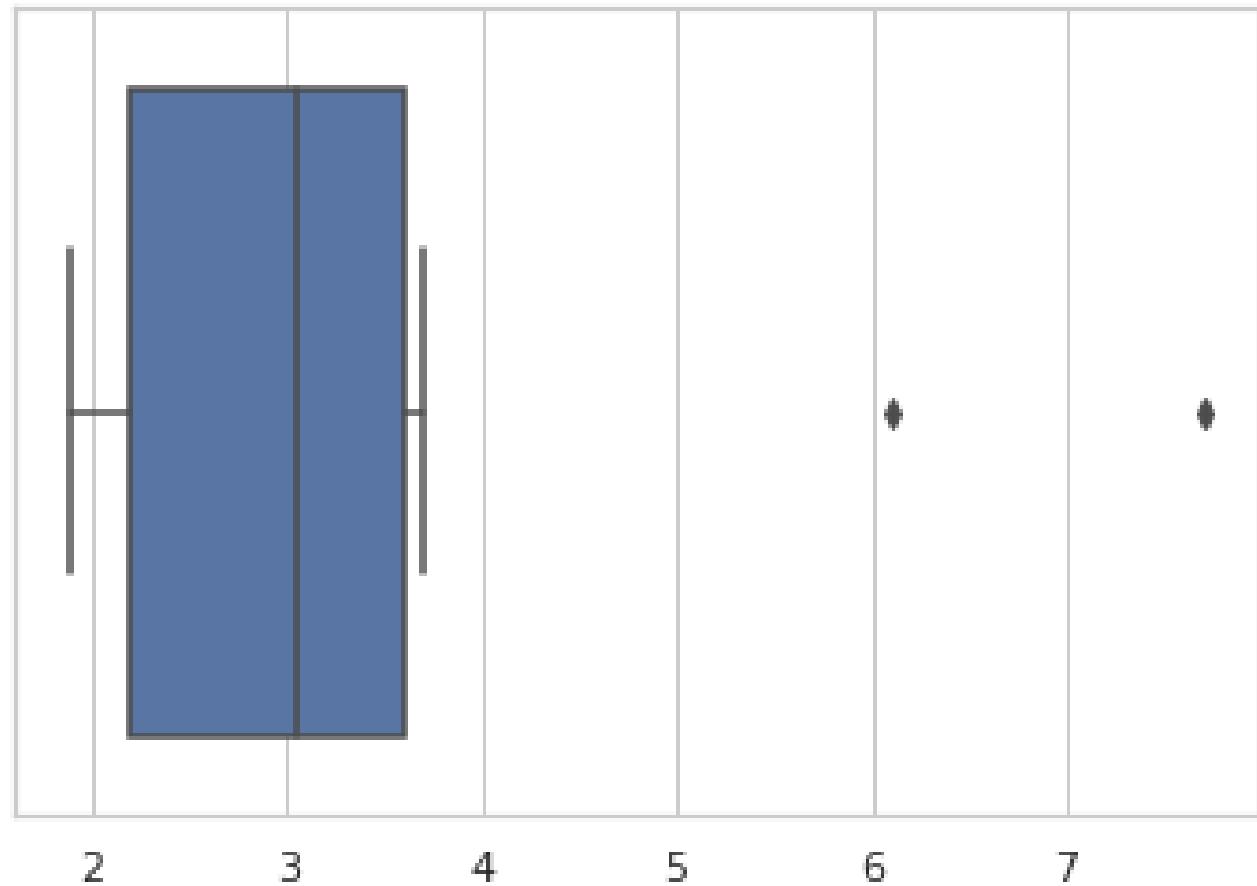
$$Q_{75\%} = Q_3 = 14$$

DIAGRAMA DE CAIXA

Valores fora destes limites são considerados outliers...



1,9 | **2,0** | **2,1** | **2,5** | **3,0** | **3,1** | **3,3** | **3,7** | **6,1** | **7,7**



TREINE

Construa o diagrama de caixa dos 40 dados a seguir

59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66
 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77

Valor mínimo: 59

Valor máximo: 77

Q2:

Distância interquartil:

$$IQD = 70 - 64,5 = 5,5$$

$$1,5 \times IQD = 8,25$$

Q1:

Q3:



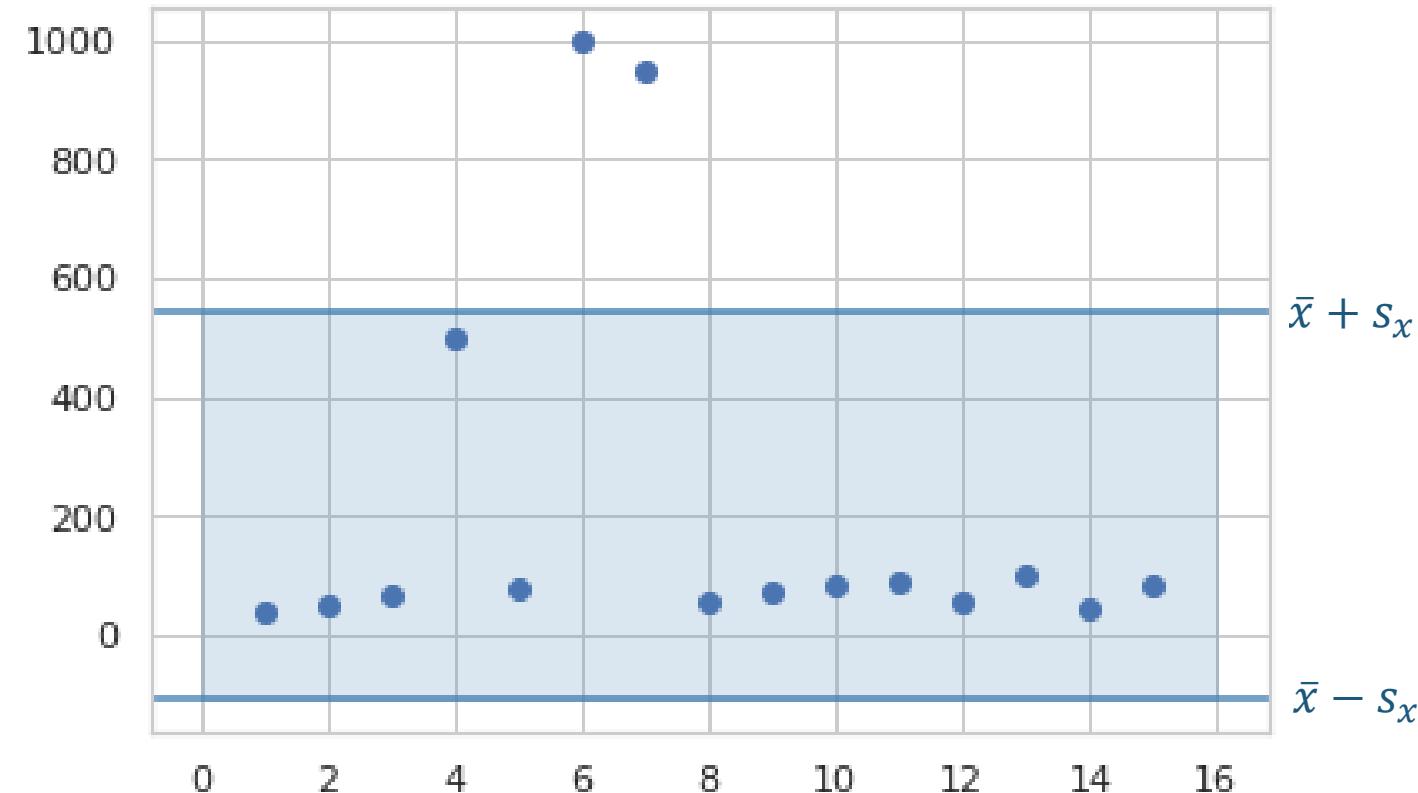
DETECÇÃO DE OUTLIER: $\bar{x} \pm s_x$

$$y = (40 \ 50 \ 70 \ 500 \ 78 \ 1000 \ 950 \ 59 \ 75 \ 85 \ 90 \ 55 \ 100 \ 44 \ 83)$$

$$n = 15$$

$$\bar{x} = 218,60$$

$$s_x = 327,01$$



DETECÇÃO DE OUTLIER: DISTÂNCIA INTERQUARTIL

$$y = (40 \ 44 \ 50 \ 55 \ 59 \ 70 \ 75 \ 78 \ 83 \ 85 \ 90 \ 100 \ 500 \ 950 \ 1000)$$

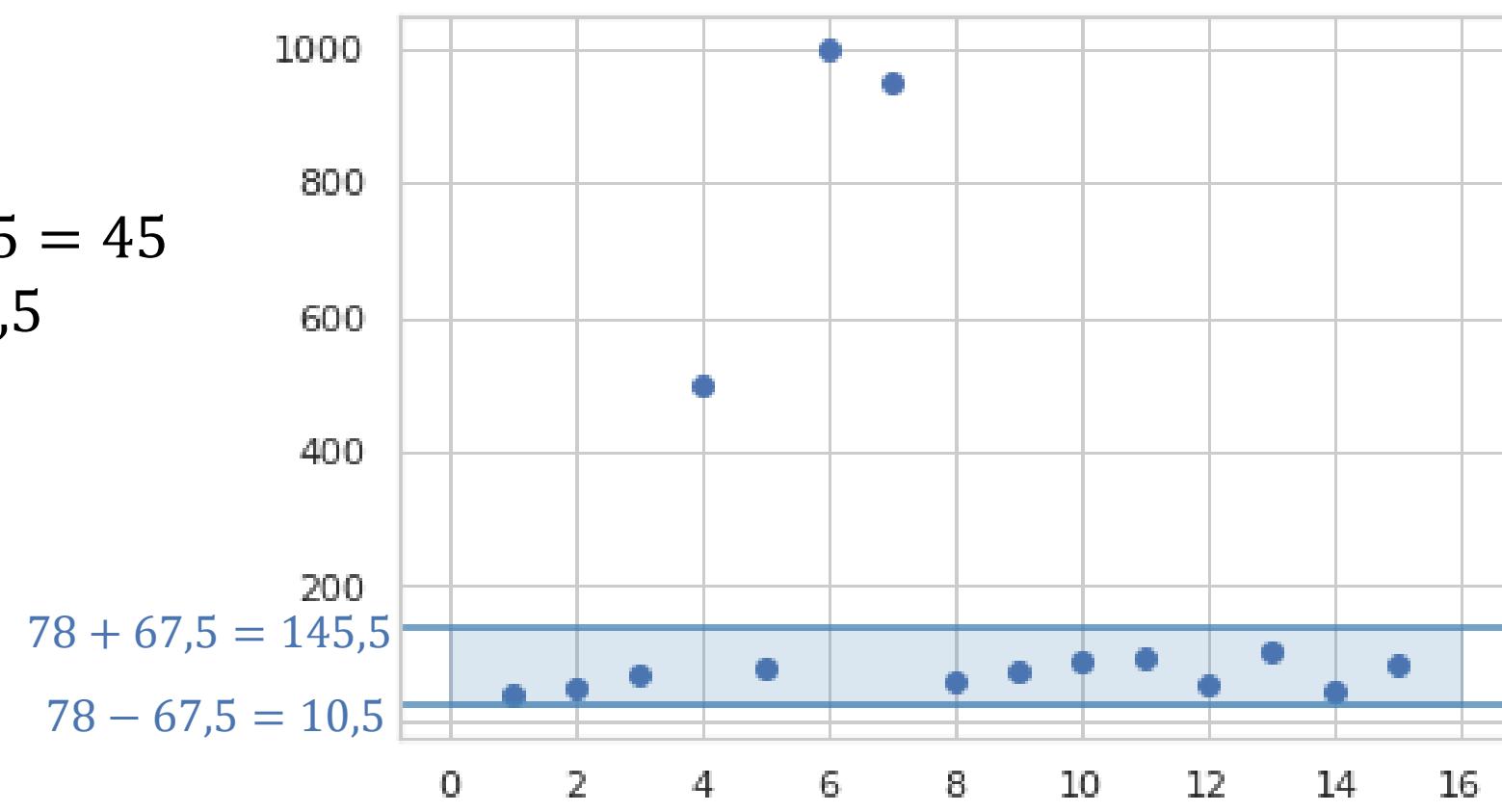
$$Q_1 = 55$$

$$Q_2 = 78$$

$$Q_3 = 100$$

$$IQD = 100 - 55 = 45$$

$$1,5 \times IQD = 67,5$$



LIÇÃO DE CASA

Para os exemplos abaixo, divida os dados em quartis.

1. Suponha a amostra de 11 elementos

72, 70, 77, 60, 67, 69, 68, 66, 65, 71, 69

2. Considere uma amostra de 6 elementos com os seguintes valores:

7,1; 7,4; 7,5; 7,7; 7,8; 7,9

3. Considere as medidas de altura de pacientes no quadro ao lado.

Altura dos pacientes			
1,59	1,79	1,68	1,80
1,58	1,60	1,69	1,73
1,87	1,68	1,85	



VARIÂNCIA, COVARIÂNCIA E CORRELAÇÃO

Variância define uma dispersão entre os dados de uma variável. A covariância e a correlação são muito úteis para entender o relacionamento entre duas variáveis.

VARIÂNCIA VS COVARIÂNCIA

Basicamente, variância mede a variação de uma única variável aleatória (como a altura de uma pessoa em uma população), enquanto covariância é uma medida de quanto duas variáveis aleatórias variam juntas (como a altura e peso de uma pessoa em uma população).

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

$$s_{xy} = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

COVARIÂNCIA VS CORRELAÇÃO

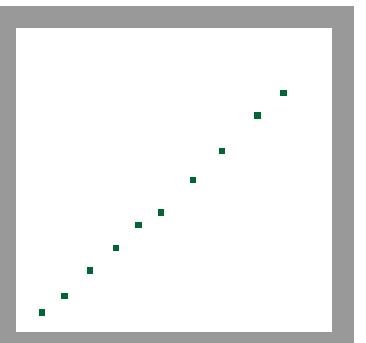
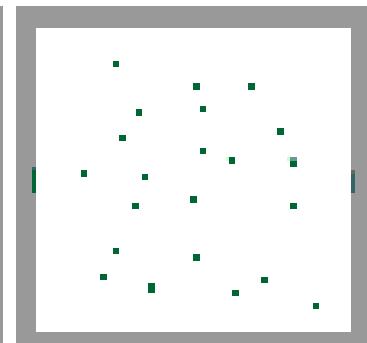
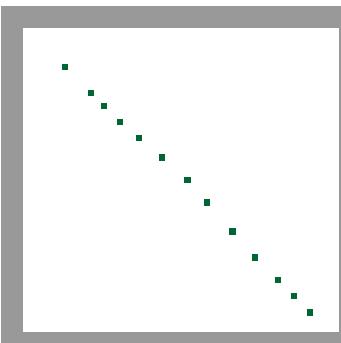
Covariância

A covariância é uma medida estatística que mostra se duas variáveis estão relacionadas, medindo como as variáveis mudam uma em relação à outra. Se ambas as variáveis variam na mesma direção tem-se **covariância positiva**; se variam em direções opostas, tem-se **covariância negativa**.

Correlação

A correlação indica a direção e o quão forte é a relação.

A correlação é a covariância normalizada pelos desvios padrão – varia entre -1 e $+1$.



COVARIÂNCIA

Covariância refere-se à medida de como duas variáveis aleatórias em um conjunto de dados serão alteradas juntas.

$$COV(x, y) = s(x, y) = s_{xy} = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Unidade da covariância é a unidade da variável x multiplicada pela unidade da variável y . Portanto, se mudarmos a unidade de variáveis, a covariância terá novo valor, apesar do sinal permanecer o mesmo:

- Covariância positiva: as duas variáveis em questão estão relacionadas positivamente e se movem na mesma direção.
- Covariância negativa: as variáveis estão inversamente relacionadas ou se movem em direções opostas.

CORRELAÇÃO

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

O coeficiente de correlação é a covariância dividida pelo desvio padrão de cada variável. É uma quantidade adimensional. Portanto, se alterarmos a unidade de x e y , o valor do coeficiente permanecerá o mesmo.

EXEMPLO

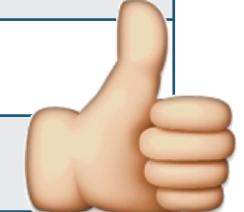
Vamos calcular covariância e correlação entre

x	2,1	2,5	3,6	4,0
y	8	10	12	14

Sequência de trabalho:

- Calcular \bar{x}, s_x ;
- Calcular \bar{y}, s_y ;
- $s_{xy} = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$;
- $r_{xy} = \frac{s_{xy}}{s_x s_y}$

Correlação r	Força da correlação
0.0 – 0.2	Fraca
0.3 – 0.6	Moderada
0.7 – 1.0	Forte



Correlation Coefficient		Dancey & Reidy (Psychology)	Quinnipiac University (Politics)	Chan YH (Medicine)
+1	-1	Perfect	Perfect	Perfect
+0.9	-0.9	Strong	Very Strong	Very Strong
+0.8	-0.8	Strong	Very Strong	Very Strong
+0.7	-0.7	Strong	Very Strong	Moderate
+0.6	-0.6	Moderate	Strong	Moderate
+0.5	-0.5	Moderate	Strong	Fair
+0.4	-0.4	Moderate	Strong	Fair
+0.3	-0.3	Weak	Moderate	Fair
+0.2	-0.2	Weak	Weak	Poor
+0.1	-0.1	Weak	Negligible	Poor
0	0	Zero	None	None

MATRIZ DE DADOS $X \in \mathbb{R}^{n \times m}$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

característica

dato

Vetor $x_j, x_k \in \mathbb{R}^n$ tal que,

$$x_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix},$$

Com média \bar{x}_j

$$x_k = \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{bmatrix},$$

Com média \bar{x}_k

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$s_{jk} = \sum_{i=1}^n \frac{(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n-1}$$



$$s_{jk} = \frac{(x_j - \mathbf{1}\bar{x}_j)^T(x_k - \mathbf{1}\bar{x}_k)}{n-1}$$

MATRIZ DE COVARIÂNCIA

Supondo que queremos a covariância entre m colunas de dados, então teremos uma matriz de covariância, $s_{ij} = s(x_i, x_j)$,

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{bmatrix}$$

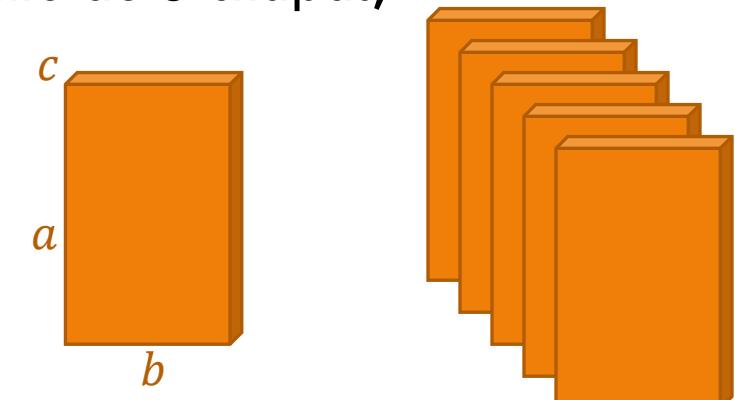
As entradas diagonais da matriz S são as **variâncias** e as entradas fora da diagonal são as **covariâncias**.

$S \in \mathbb{R}^{m \times m}$ é uma matriz quadrada e simétrica.

EXEMPLO

Dado o conjunto de observações das três dimensões de um conjunto de 5 chapas,

$$X = \begin{bmatrix} 4,1 & 2,0 & 0,60 \\ 4,2 & 2,1 & 0,59 \\ 3,9 & 2,0 & 0,58 \\ 4,3 & 2,1 & 0,62 \\ 4,1 & 2,2 & 0,63 \end{bmatrix}$$



Calcule o vetor \bar{x} e a matriz de covariância

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix}$$

$$X = \begin{bmatrix} 4,1 & 2,0 & 0,60 \\ 4,2 & 2,1 & 0,59 \\ 3,9 & 2,0 & 0,58 \\ 4,3 & 2,1 & 0,62 \\ 4,1 & 2,2 & 0,63 \end{bmatrix}$$

$$\bar{x}_1 = \frac{4,1+4,2+3,9+4,3+4,1}{5} = 4,12$$

Analogamente,

$$\bar{x}_2 = 2,08$$

$$\bar{x}_3 = 0,604$$

$$\bar{x} = [4,12 \quad 2,08 \quad 0,604]$$

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix}$$

$$s_{jk} = \frac{(\mathbf{x}_j - \mathbf{1}\bar{x}_j)^T (\mathbf{x}_k - \mathbf{1}\bar{x}_k)}{n-1}$$

$$s_{23} = \frac{(\mathbf{x}_2 - \mathbf{1}\bar{x}_2)^T (\mathbf{x}_3 - \mathbf{1}\bar{x}_3)}{4}$$

$$x_2 \quad x_3$$

$$X = \begin{bmatrix} 4,1 & 2,0 & 0,60 \\ 4,2 & 2,1 & 0,59 \\ 3,9 & 2,0 & 0,58 \\ 4,3 & 2,1 & 0,62 \\ 4,1 & 2,2 & 0,63 \end{bmatrix}$$

$$\mathbf{x}_2 - \mathbf{1}\bar{x}_2 = \begin{bmatrix} 2,0 - 2,08 \\ 2,1 - 2,08 \\ 2,0 - 2,08 \\ 2,1 - 2,08 \\ 2,2 - 2,08 \end{bmatrix} = \begin{bmatrix} -0,08 \\ 0,02 \\ -0,08 \\ 0,02 \\ 0,12 \end{bmatrix}$$

$$\mathbf{x}_3 - \mathbf{1}\bar{x}_3 = \begin{bmatrix} 0,60 - 0,604 \\ 0,59 - 0,604 \\ 0,58 - 0,604 \\ 0,62 - 0,604 \\ 0,63 - 0,604 \end{bmatrix} = \begin{bmatrix} -0,004 \\ -0,014 \\ -0,024 \\ 0,016 \\ 0,026 \end{bmatrix}$$

$$\frac{1}{4} [-0,08 \quad 0,02 \quad -0,08 \quad 0,02 \quad 0,12] \begin{bmatrix} -0,004 \\ -0,014 \\ -0,024 \\ 0,016 \\ 0,026 \end{bmatrix} = 0,00135$$

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & 0,0014 \\ s_{31} & 0,0014 & s_{33} \end{bmatrix} \rightarrow S = \begin{bmatrix} 0,0220 & 0,0055 & 0,0016 \\ 0,0055 & 0,0070 & 0,0014 \\ 0,0016 & 0,0014 & 0,0004 \end{bmatrix}$$

```

M=np.array([[4.1, 2.0, 0.60],
            [4.2, 2.1, 0.59],
            [3.9, 2.0, 0.58],
            [4.3, 2.1, 0.62],
            [4.1, 2.2, 0.63]])

print('Matriz M com os dados')
print(M.T)
cov_M = np.cov(M.T)
print('Matriz de covariância dos dados')
print(cov_M)

```

```

Matriz M com os dados
[[4.1 4.2 3.9 4.3 4.1 ]
 [2. 2.1 2. 2.1 2.2 ]
 [0.6 0.59 0.58 0.62 0.63]]
Matriz de covariância dos dados
[[0.022 0.0055 0.00165]
 [0.0055 0.007 0.00135]
 [0.00165 0.00135 0.00043]]

```

$$S = \begin{bmatrix} 0,0220 & 0,0055 & 0,0016 \\ 0,0055 & 0,0070 & 0,0014 \\ 0,0016 & 0,0014 & 0,0004 \end{bmatrix}$$

MATRIZ DE CORRELAÇÃO

Supondo que queremos a correlação entre m dados, então teremos uma matriz de correlação, $r_{ij} = r(x_i, x_j)$,

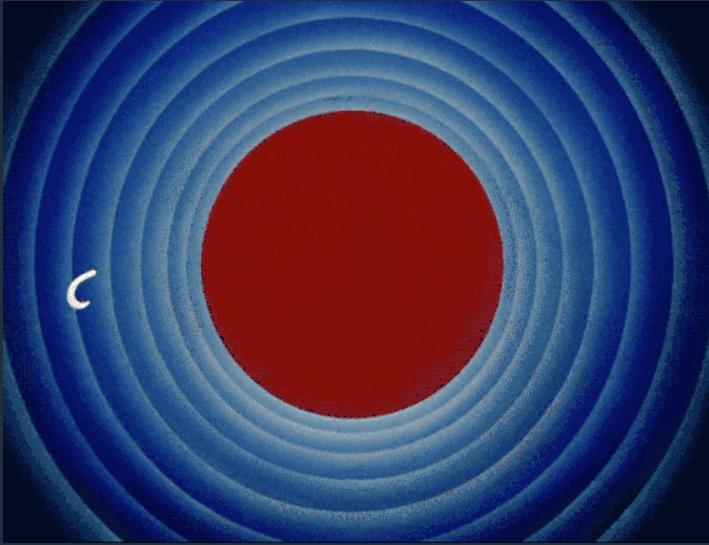
$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{bmatrix}$$

$R \in \mathbb{R}^{m \times m}$ é uma matriz quadrada e simétrica.

TAREFA

Aluno	Python	Estatística	Arte
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

Calcule a matriz de correlação dos dados acima e discuta os resultados.



ACABOU

Próxima aula:
Probabilidade