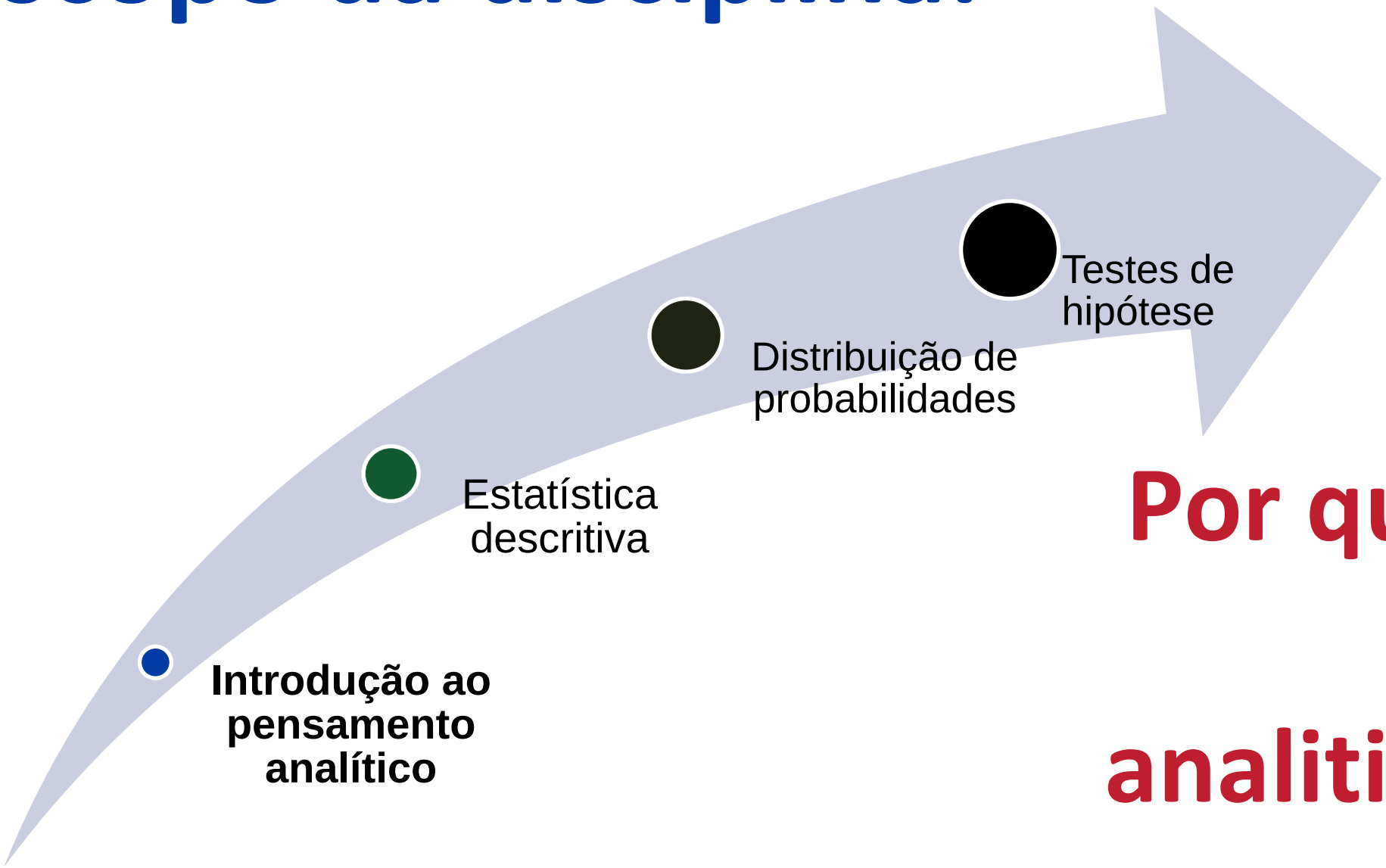
Desenvolvimento do pensamento analítico/estatístico

INTELIGÊNCIA ARTIFICIAL NA SAÚDE

Prof. Jonathan Gil Müller

Escopo da disciplina:



Por que e como pensar analiticamente?

Prof. Jonathan Gil Müller <jgmuller@furb.br>

O que é pensamento analítico?

Informalmente, dizemos que é o **pensamento realizado para analisar algo**.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Por que analisar?

Para **extrair informações**, **tirar conclusões**, desenvolver **novos conhecimentos**, entre outros objetivos

Prof. Jonathan Gil Müller <jgmuller@furb.br>

O que se analisa em Inteligência Artificial (IA)?

Dados!

Prof. Jonathan Gil Müller <jgmuller@furb.br>

O que é um conjunto de dados?

Prof. Jonathan Gil Müller <jgmuller@furb.br>

O que é um conjunto de dados?

Variáveis

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	style
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	red
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	red
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	red
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	red
5	7.4	0.66	0.00	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5	red
6	7.9	0.60	0.06	1.6	0.069	15.0	59.0	0.9964	3.30	0.46	9.4	5	red
7	7.3	0.65	0.00	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7	red
8	7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7	red
9	7.5	0.50	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.80	10.5	5	red

Indivíduos ou casos

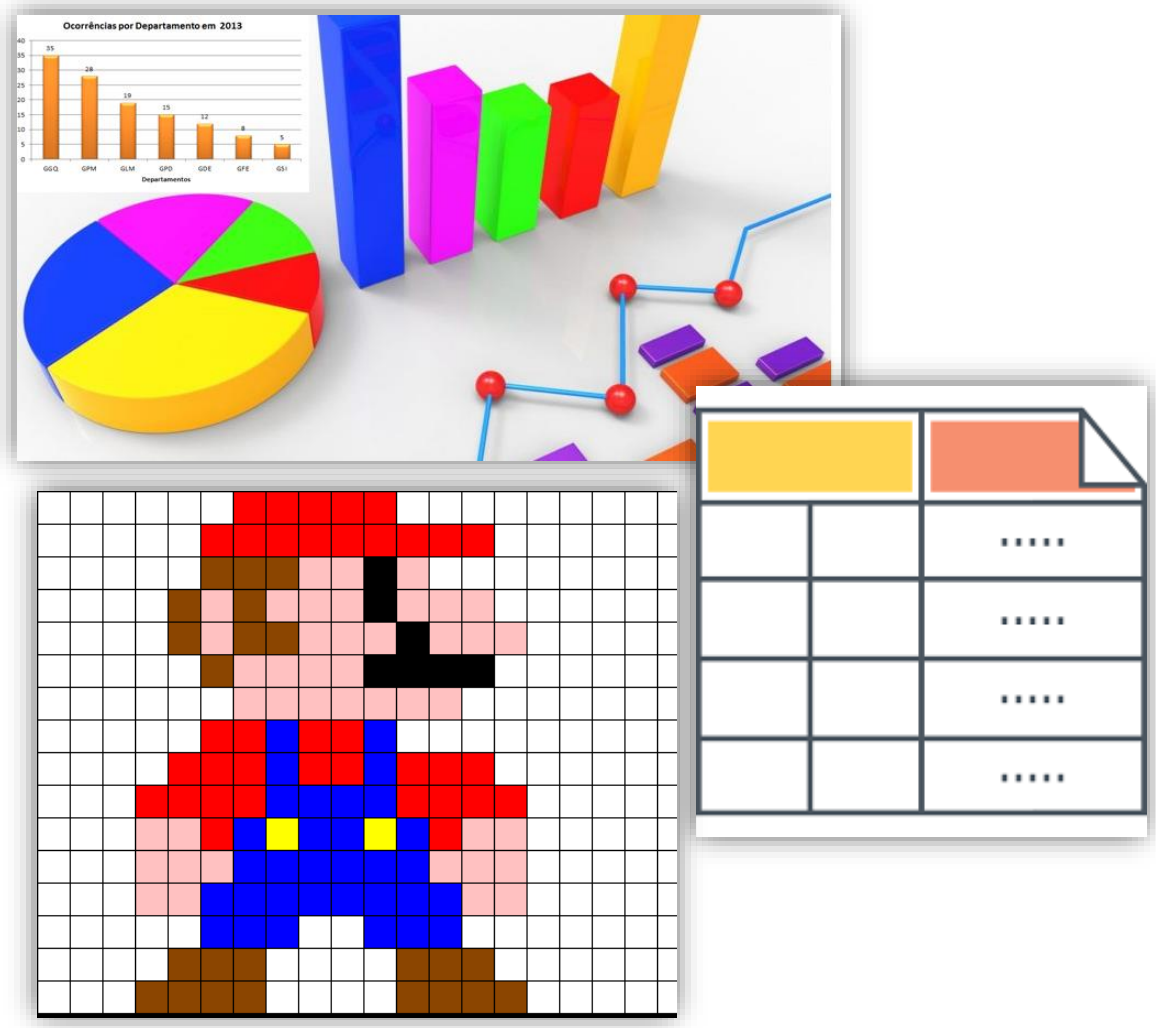
Dados

Prof. Jonathan Gil Müller <jgmuller@furb.br>

O que é um conjunto de dados?

- Os dados podem ser **apresentados de maneiras diferentes**, as vezes parece até que estão “disfarçados” (tabelas, gráficos, imagens, matrizes etc.).

- Na verdade, eles já **passaram por algumas etapas de processamento/pré-processamento**, ou seja, alguém já os analisou e realizou algum tratamento matemático e/ou computacional neles.



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Áreas da estatística

- Estatística
  - Estatística descritiva
  - Probabilidade
  - Inferência estatística

Nesta disciplina, discutiremos os alguns **conceitos básicos** da estatística que **fundamentam o pensamento analítico** e, consequentemente, a **análise de dados**.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Áreas da estatística

É a **etapa inicial** da análise de dados e pode ser entendida com um conjunto de **parâmetros** que podemos utilizar para **organizar, resumir e conhecer** nosso conjunto de dados. Dependendo da situação, a estatística descritiva também pode auxiliar na **coleta de dados**.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Áreas da estatística

Falaremos mais especificamente da **distribuição de probabilidade** de um conjunto de dados. Este tópico nos ajudará a **conhecer como os dados se distribuem** dentro de uma escala e também a **lidar com fenômenos aleatórios**.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

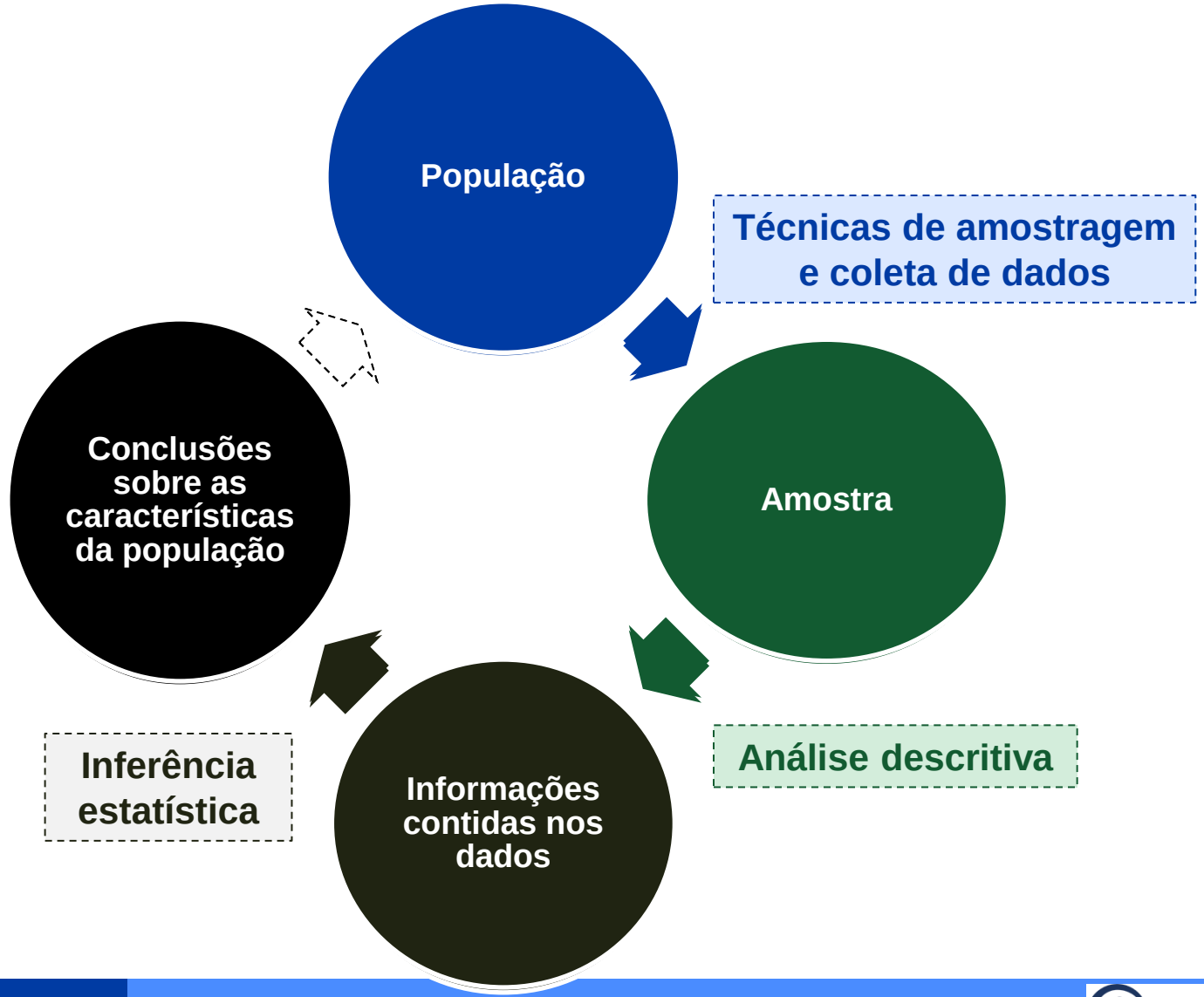
Áreas da estatística

Novos conhecimentos e a solução de alguns problemas podem ser alcançados através de uma **hipótese bem sucedida** (aceita/corroborada). Como podemos verificar se uma hipótese será corroborada ou não? Uma alternativa é por meio dos **testes de hipótese**.

Informalmente, podemos entendê-la como uma "caixinha" aonde estão guardadas as ferramentas que utilizamos para **extrapolar informações e conclusões** do nosso conjunto de dados.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

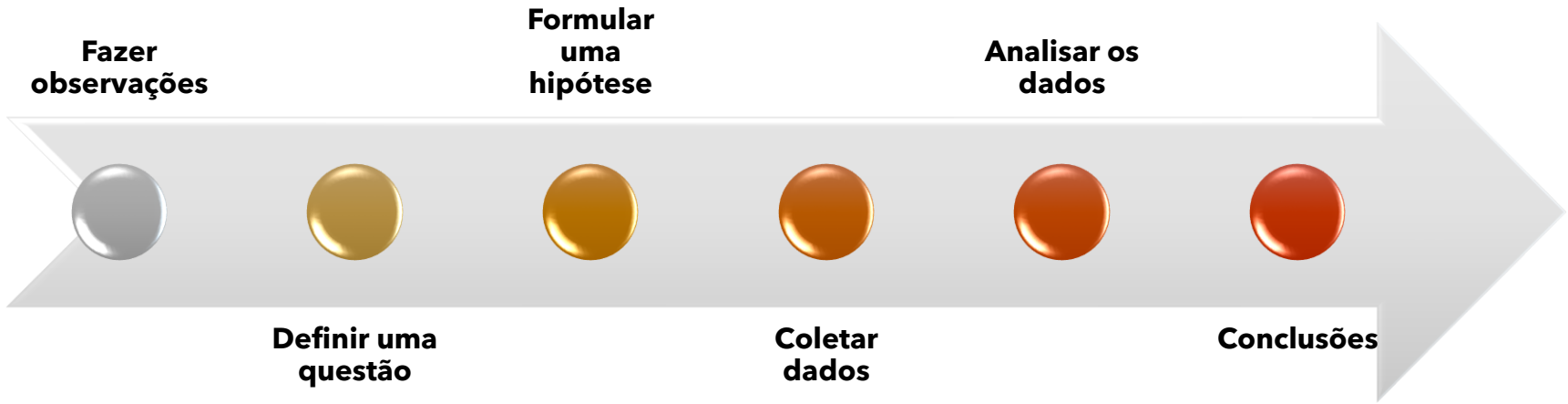
Etapas da análise estatística



Prof. Jonathan Gil Müller <jgmuller@furb.br>

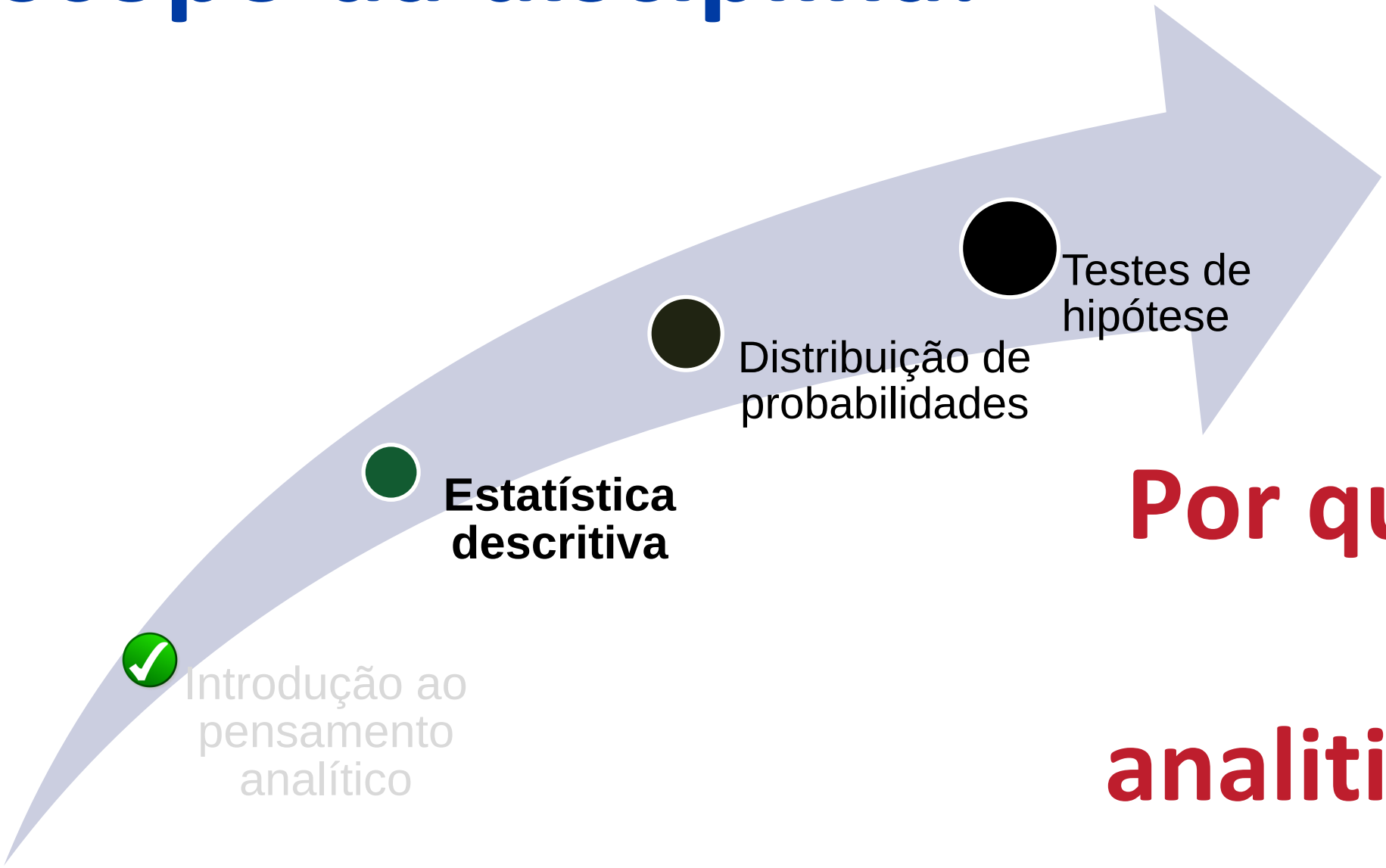
Enfim... O que é estatística?

- A estatística é um **conjunto de técnicas** que permite, de forma sistemática, **organizar, descrever, analisar e interpretar** dados.
- Sendo assim, a estatística disponibiliza **ferramentas para desenvolver o pensamento analítico** capaz de obter conclusões de um conjunto de dados.



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Escopo da disciplina:



Por que e como pensar analiticamente?

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Estatística descritiva

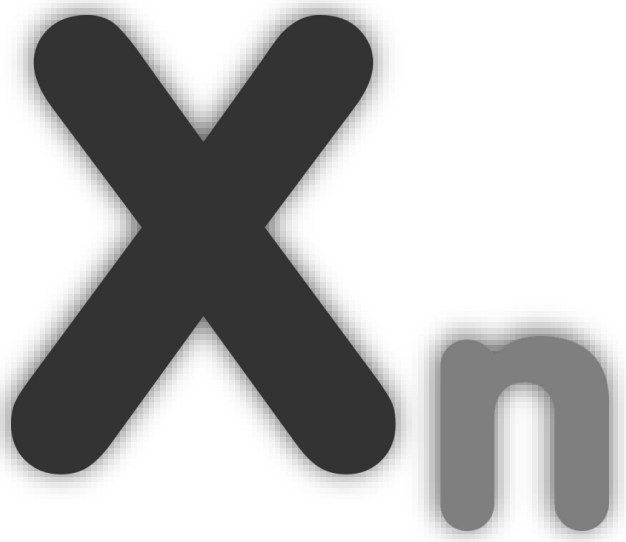
Após coletar e tabular as informações de um banco de dados, **qual o primeiro passo para prosseguir com a análise**?

Resumir os dados. Para isso a estatística descritiva tem alguns **parâmetros** que podem nos auxiliar. Vamos falar sobre eles e compreender o conceito de cada um.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Variáveis

Antes de falarmos dos parâmetros que compõe a estatística descritiva, vamos falar brevemente do **conceito de variável**.



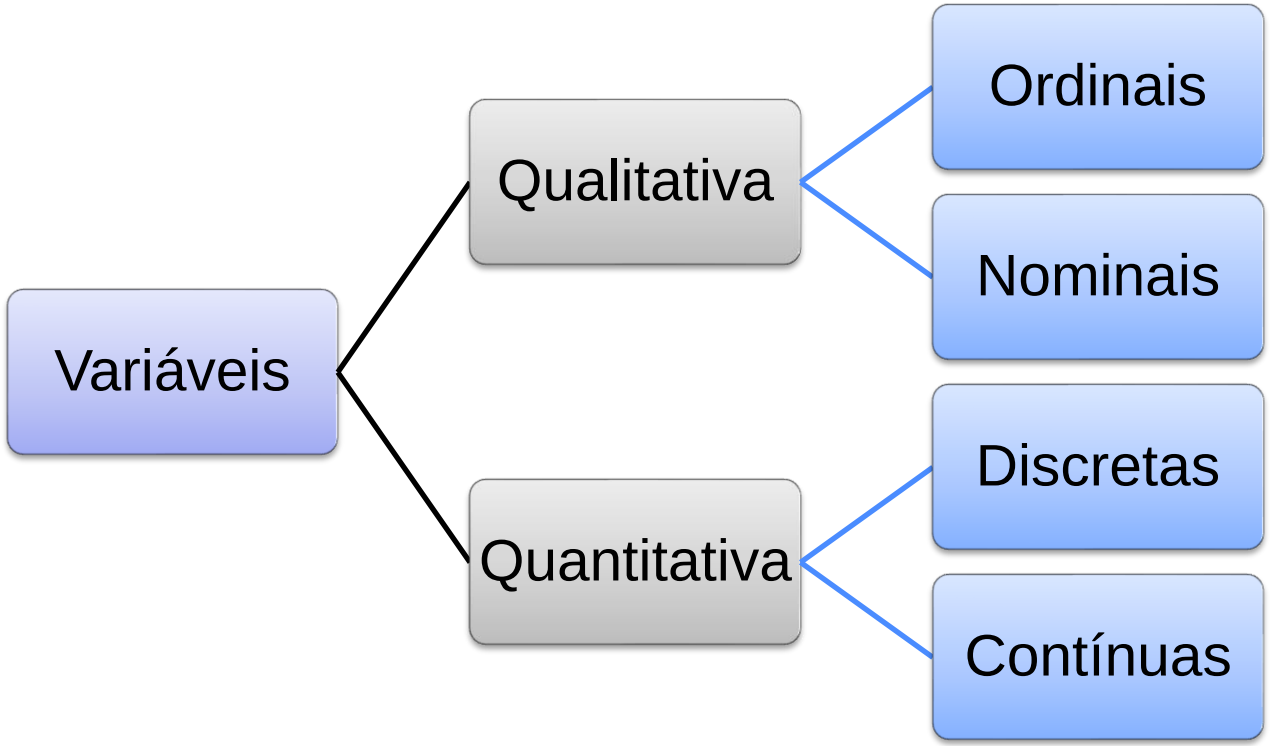
Uma variável é utilizada para **armazenar uma determinada característica** (ou a variação dessa característica) de uma população/amostra.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Variáveis

Quais são os tipos de variáveis?

Elas são classificadas de acordo com o **tipo de informação** que ela está armazenando.



Prof. Jonathan Gil Müller <jgmuller@furb.br>

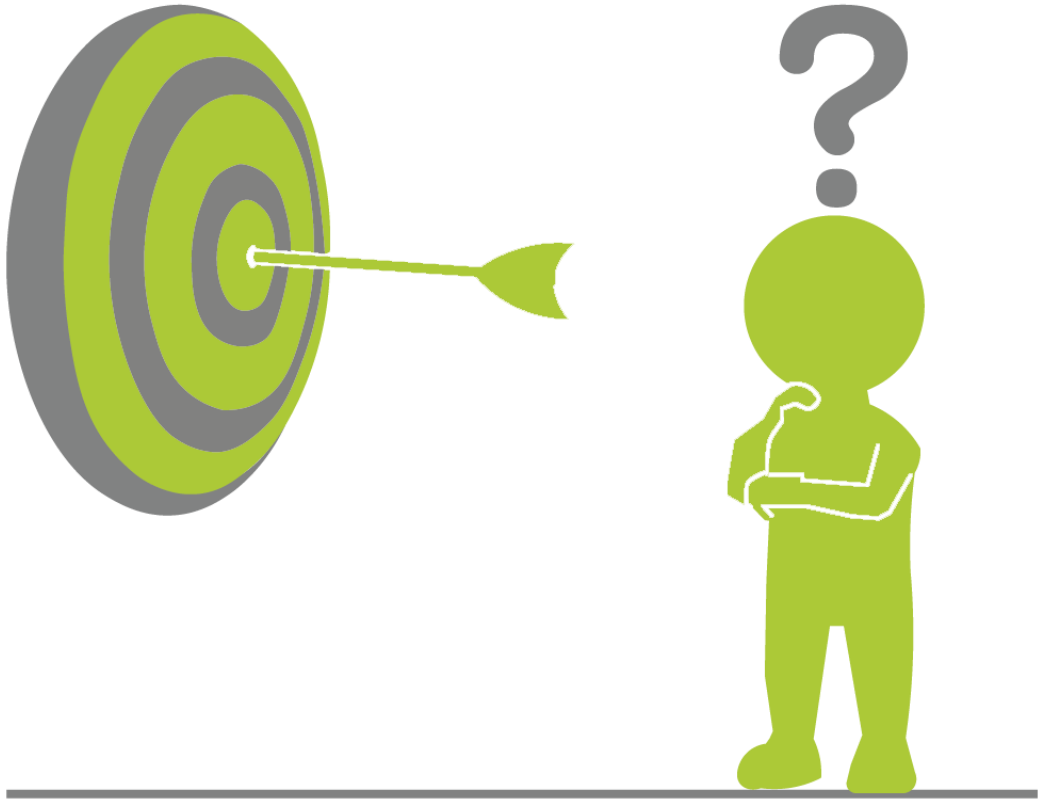
Estatística descritiva

Um conjunto de dados pode ser composto por uma ou mais variáveis (análise univariada, bivariada e multivariada) e a informação de cada variável pode ser resumida por:

- **medidas de posição:** são valores que representa uma tendência central dos dados;
- **medidas de dispersão:** são valores que indicam como os dados estão dispersos em relação ao valor central do estudo, ou seja, resume a variabilidade dos dados daquela variável.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição



Qual a finalidade das medidas de posição?

Sua finalidade é **determinar um valor central** (ou um valor típico) que represente o nosso conjunto de dados.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

MÉDIA:

A medida de tendência central mais comum é a média. É obtida pela soma das observações amostrais, dividida pelo número total de observações. Por exemplo, sejam as idades de $n = 8$ pessoas, em anos completos:

38, 40, 49, 67, 33, 57, 54 e 64.

A média amostral, denotada por $\bar{x}$ (leia "xis barra"), é dada por

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

Assim,

$$\bar{x} = \frac{38 + 40 + 49 + 67 + 33 + 57 + 54 + 64}{8} = \frac{402}{8} = 50,25 \text{ anos.}$$

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

MÉDIA:

Observe que a média é expressa na mesma unidade de medida da variável em questão.

Como interpretamos uma média de $\bar{x} = 50{,}25$ anos? Em primeiro lugar, sendo a média uma medida de tendência central, podemos dizer que as idades das $n = 8$ pessoas de nossa amostra flutuam em torno de 50,25 anos. Observe a Figura 3.1.

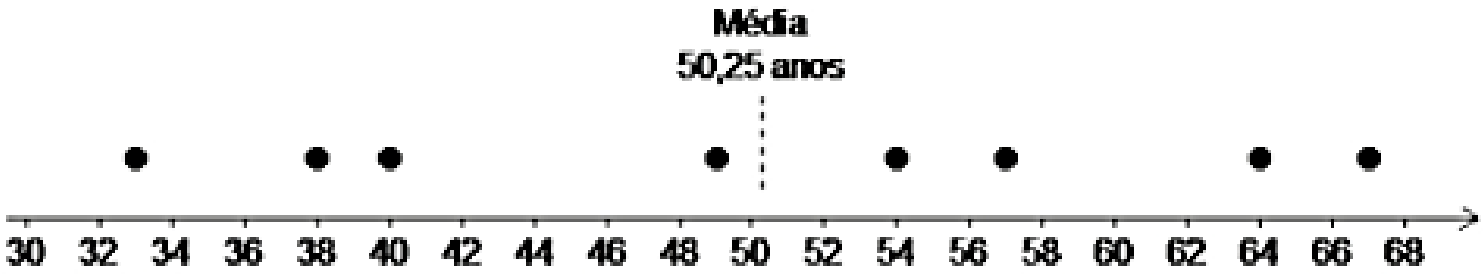


Figura 3.1 Média das idades de 8 pessoas.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

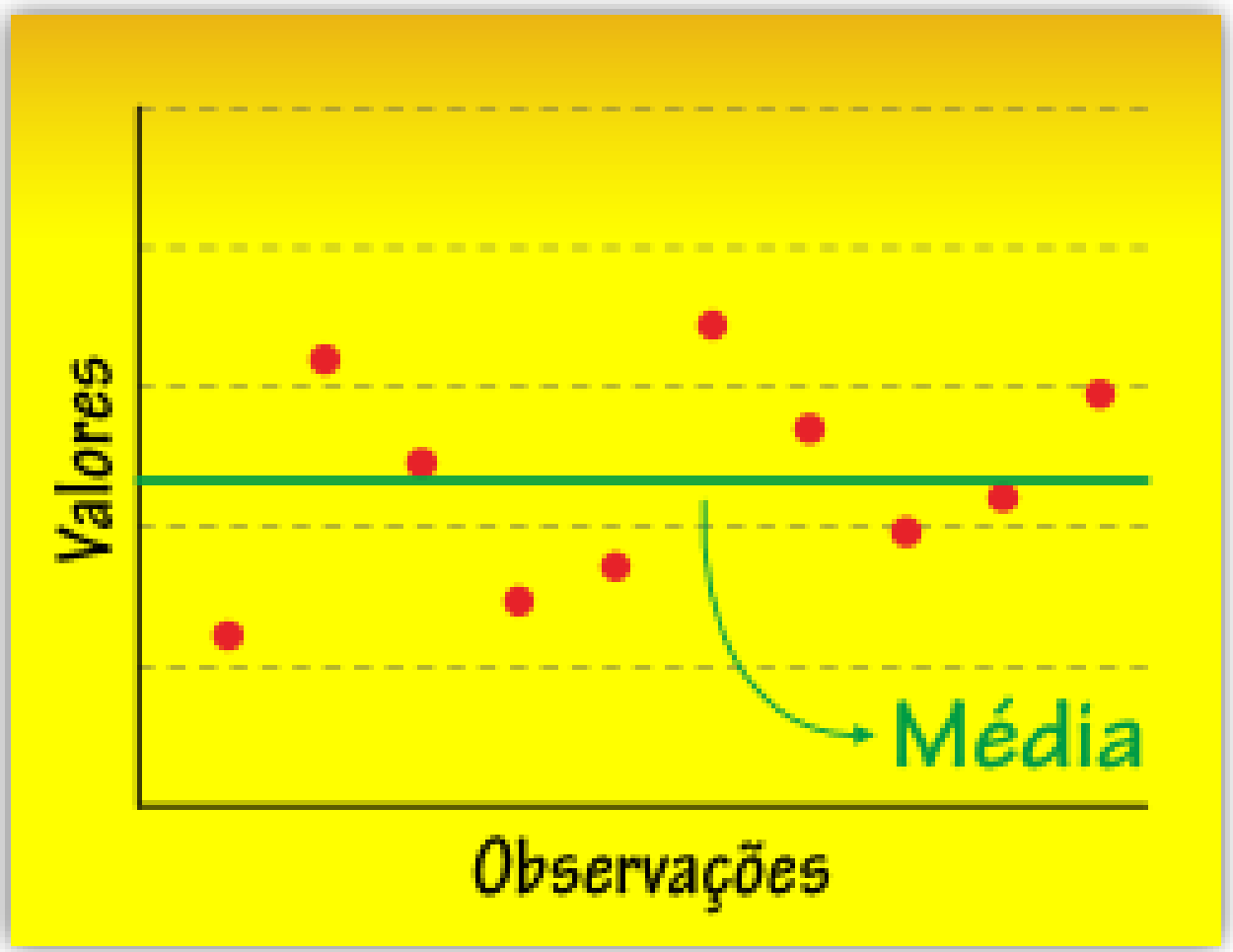
MÉDIA:

Em segundo lugar, lembre-se de que a média é uma medida-resumo, ou seja, ela objetiva **sintetizar em um único valor todas as nossas observações amostrais**. Em outras palavras, a idade de 50,25 anos é um valor que busca representar as idades de todas as n = 8 pessoas. Entretanto, observe que **a média é um resumo incompleto** de nosso conjunto de dados, dado que ela não informa o tamanho da dispersão de nossos dados a seu redor. Para essa finalidade, existe o desvio padrão, que discutiremos posteriormente.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

MÉDIA:



Medidas de posição

MEDIANA:

Assim como a média, a mediana também é uma medida de tendência central. Entretanto, ela é obtida de uma maneira diferente. Se as observações são ordenadas da menor até a maior, metade dos valores é maior ou igual à mediana, enquanto a outra metade é menor ou igual a ela. Por exemplo, sejam as estaturas de $n = 13$ mulheres, em metros, exibidas a seguir.

1,59	1,51	1,63	1,58	1,52	1,52	1,64
1,66	1,53	1,50	1,60	1,56	1,64	

O primeiro passo consiste em ordenar as observações, da menor para a maior:

1,50 1,51 1,52 1,52 1,53 1,56 1,58 1,59 1,60 1,63 1,64 1,64 1,66

Em segundo lugar, selecionamos o "número do meio", que é 1,58 m:

1,50 1,51 1,52 1,52 1,53 1,56 <u>1,58</u> 1,59 1,60 1,63 1,64 1,64 1,66

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

Esse número é a mediana. A Figura 3.2 ilustra a posição da mediana na série de observações. Perceba que há um mesmo número de observações antes e após a mediana.

MEDIANA:

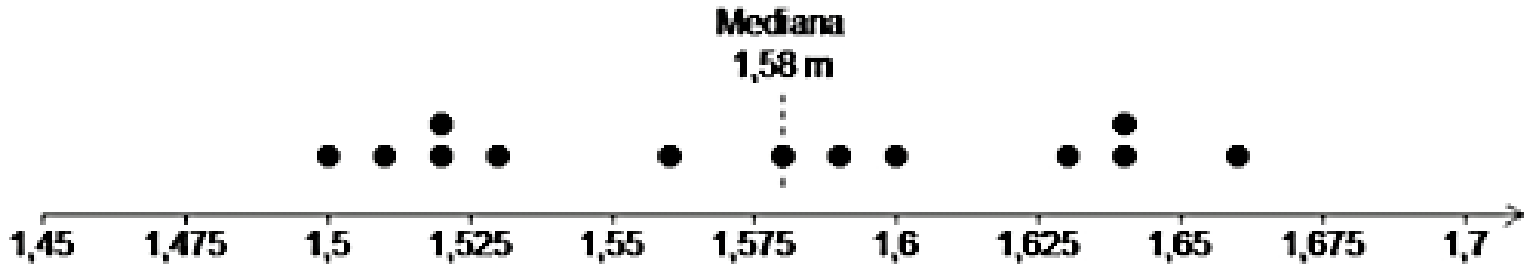
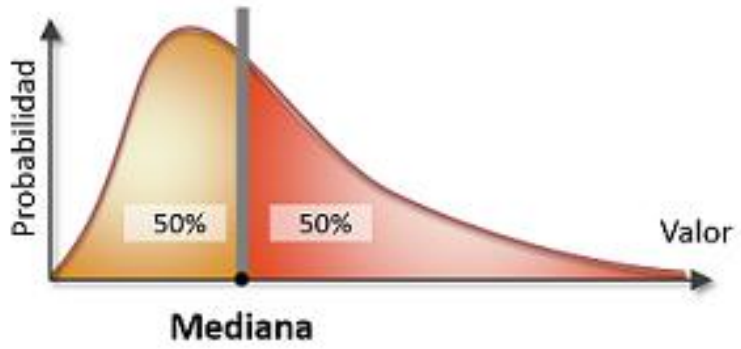


Figura 3.2 Mediana das estaturas (em m) de 13 mulheres.



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

MEDIANA:

Se substituirmos a mulher de maior estatura (1,66 m) por outra ainda mais alta (digamos, de 1,90 m), o que aconteceria com a mediana? Seu valor se modificaria?

O “número do meio” continuaria sendo 1,58 m. Esta é uma característica importante da mediana: ela não é sensível a valores atípicos de nosso conjunto de dados (entendemos por valor atípico um número bastante grande ou pequeno em relação aos demais).

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

MEDIANA:

E o que aconteceria com a média amostral ao substituirmos a mulher de 1,66 m de estatura pela outra de 1,90 m?

A média amostral ficaria um pouco maior. Dado que a média costuma ser sensível ao efeito de valores atípicos, a mediana é muitas vezes utilizada quando esses valores demasiadamente grandes ou pequenos estão presentes em nossos dados.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

MEDIANA:

No exemplo anterior, encontramos a mediana considerando um número ímpar de observações. E se tivermos um número par de observações? Sejam novamente as idades de $n = 8$ pessoas, em anos completos:

38, 40, 49, 67, 33, 57, 54 e 64

Depois de ordenarmos as observações, da menor até a maior, observamos que os "números do meio" são 49 e 54.

33 38 40 <u>49</u> <u>54</u> 57 64 67

A mediana é o valor central entre 49 e 54, dado por:

$$\frac{49 + 54}{2} = 51,5 \text{ anos.}$$

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição

MODA:

A moda é a observação que ocorre com maior frequência no conjunto de dados.

Imagine que em uma loja de sapatos femininos foram vendidos 20 pares de sapatos em um único dia. Os sapatos vendidos tinham estas numerações:

34	34	35	35	36	36	36	36	36	36
37	37	37	37	38	38	38	39	39	39

A numeração mais frequente entre os sapatos vendidos é 36. Essa informação é de grande utilidade ao gerente da loja, pois indica que ele não pode deixar de ter sapatos número 36 em seu estoque. Perceba que nesse exemplo a mediana ou a média não teria a mesma utilidade.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

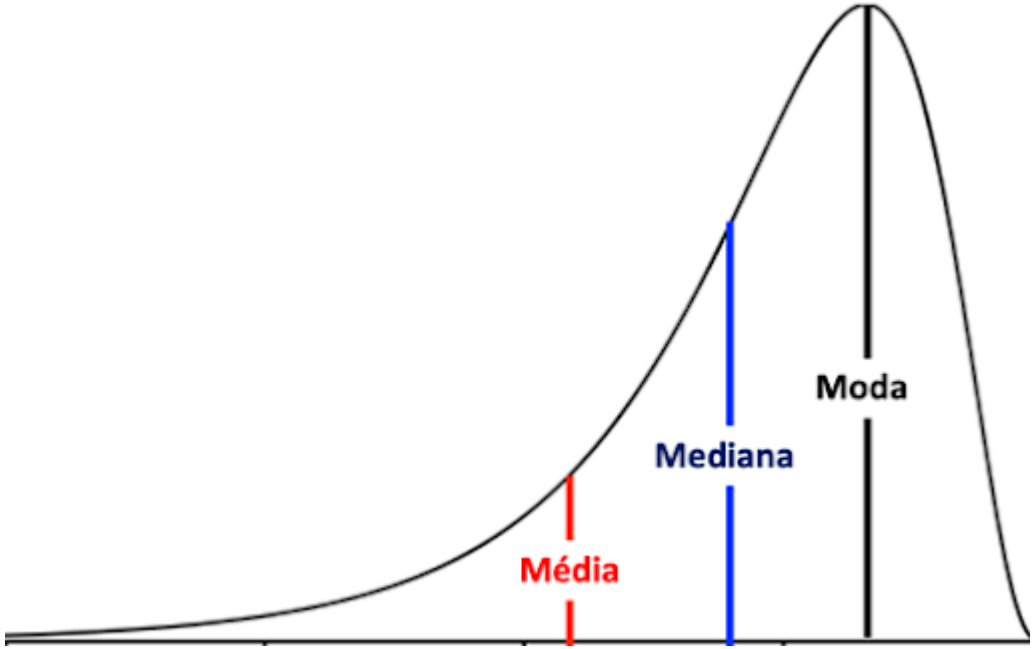
Medidas de posição

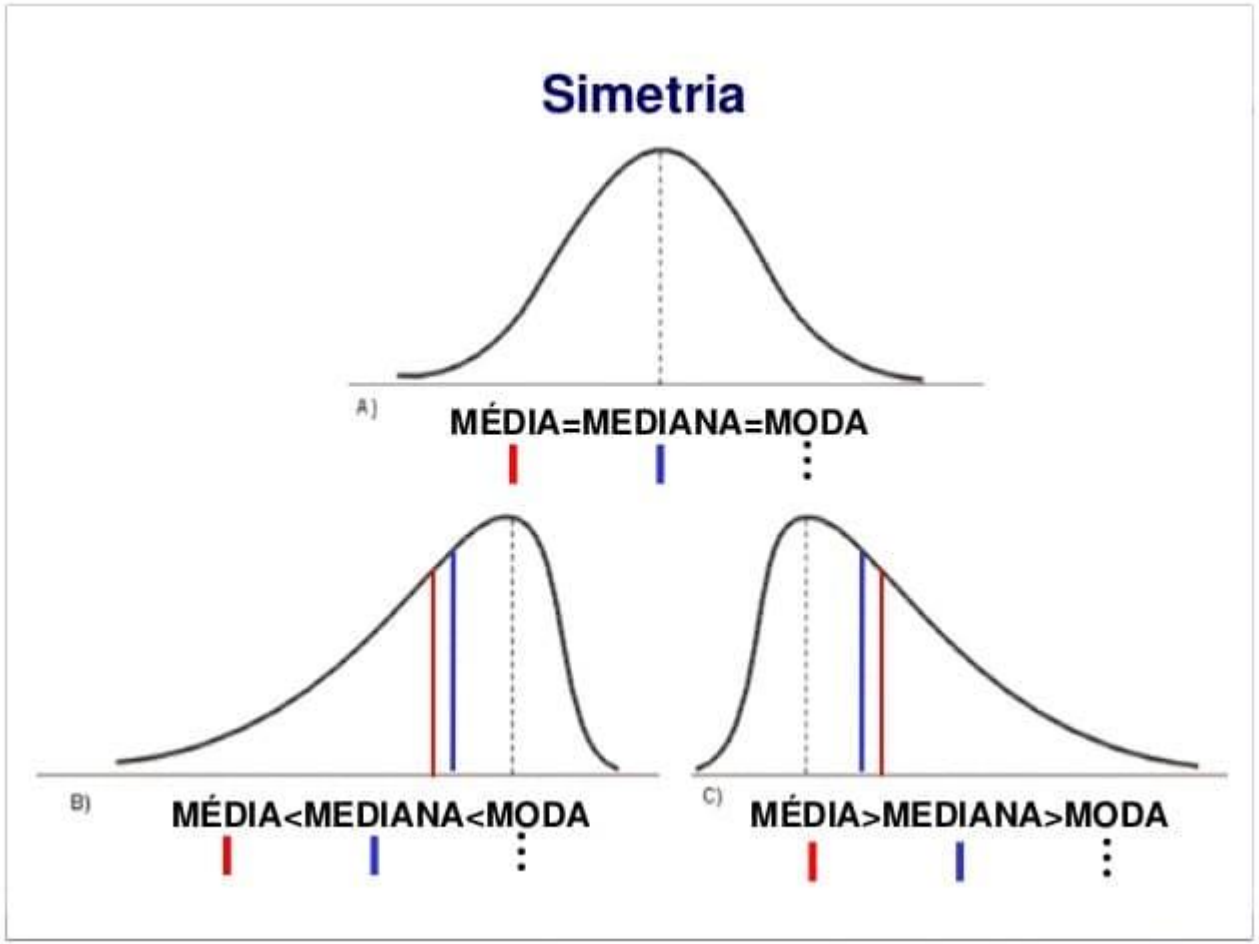
MODA:

É importante não confundir moda com maioria. A moda é a observação mais frequente, mas isso não implica necessariamente que a moda corresponde à maioria das observações. Nesse exemplo, a moda é 36, mas observamos que não é válido dizer que a maioria dos sapatos vendidos tem numeração 36. Seis sapatos número 36 em um total de 20 correspondem a 30%. A maioria corresponderia a 50% dos sapatos mais um, ou seja, 11 sapatos em 20.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Relação entre medidas de posição



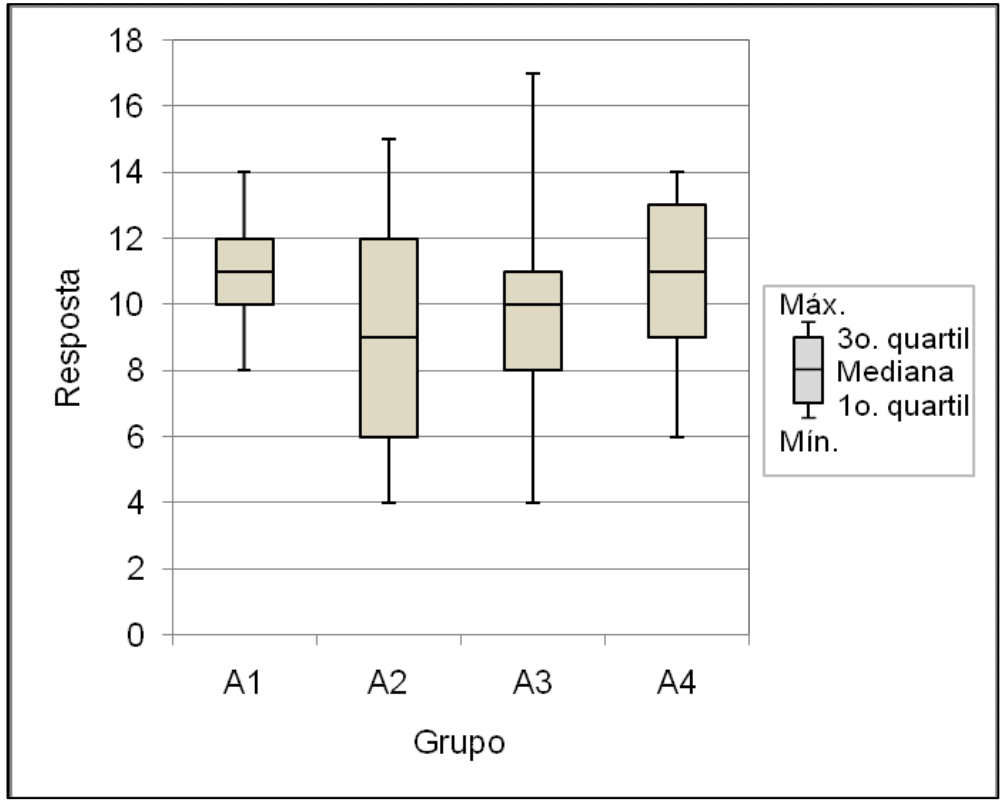


Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição e dispersão

Tem situações onde é relevante conhecer outros aspectos referente ao conjunto de valores, além de um valor central.

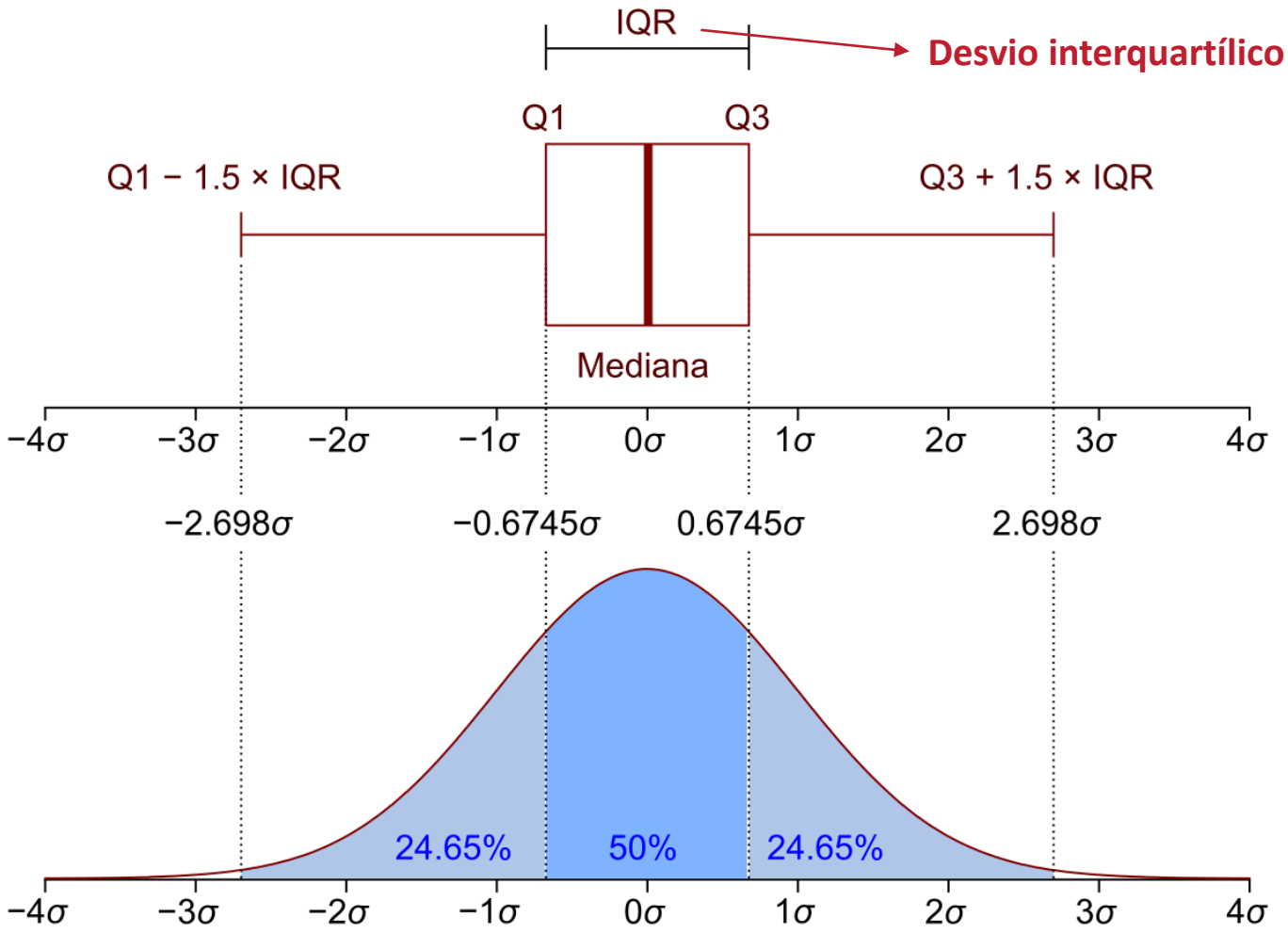
QUARTIS e EXTREMOS:



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição e dispersão

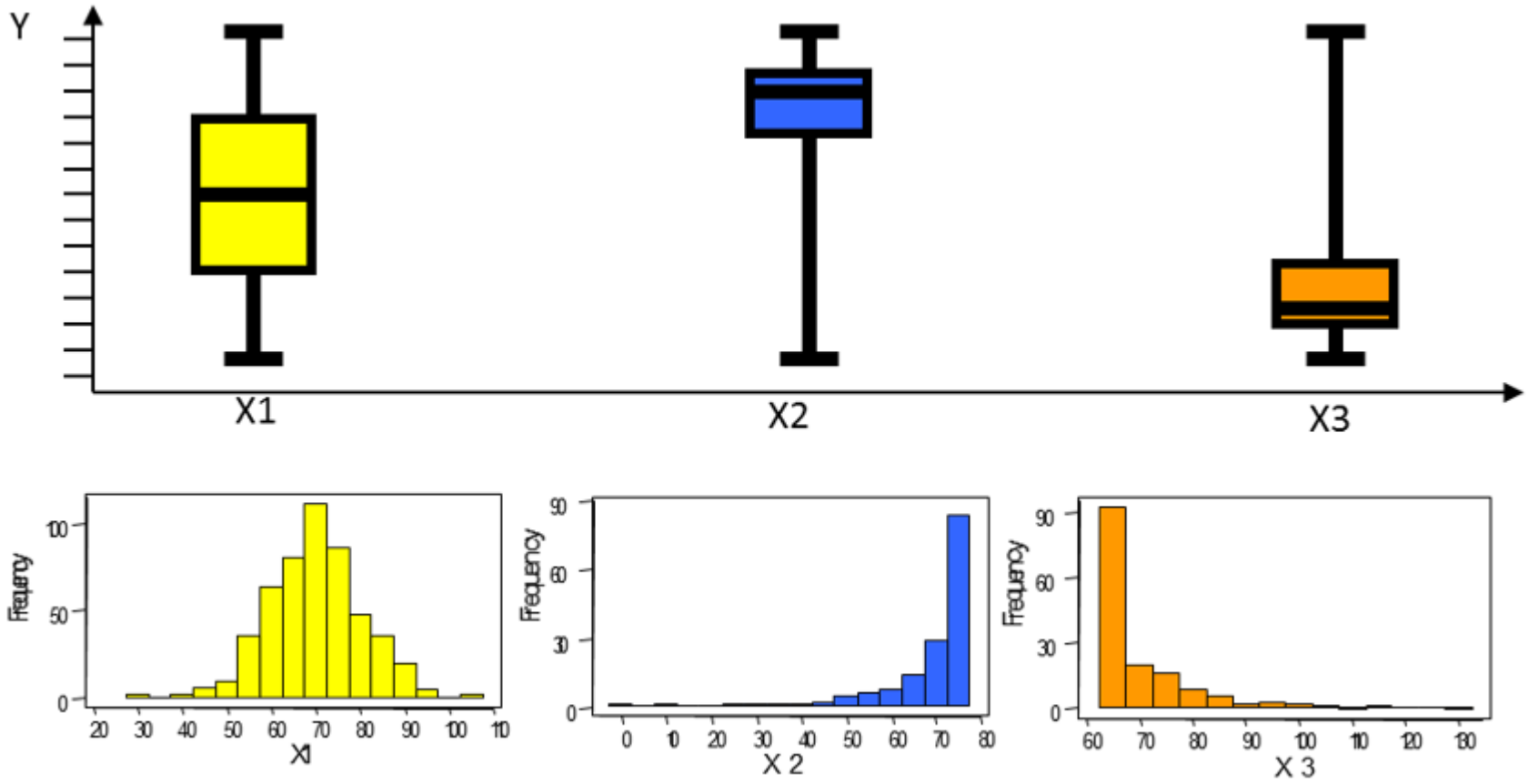
QUARTIS e EXTREMOS:



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição e dispersão

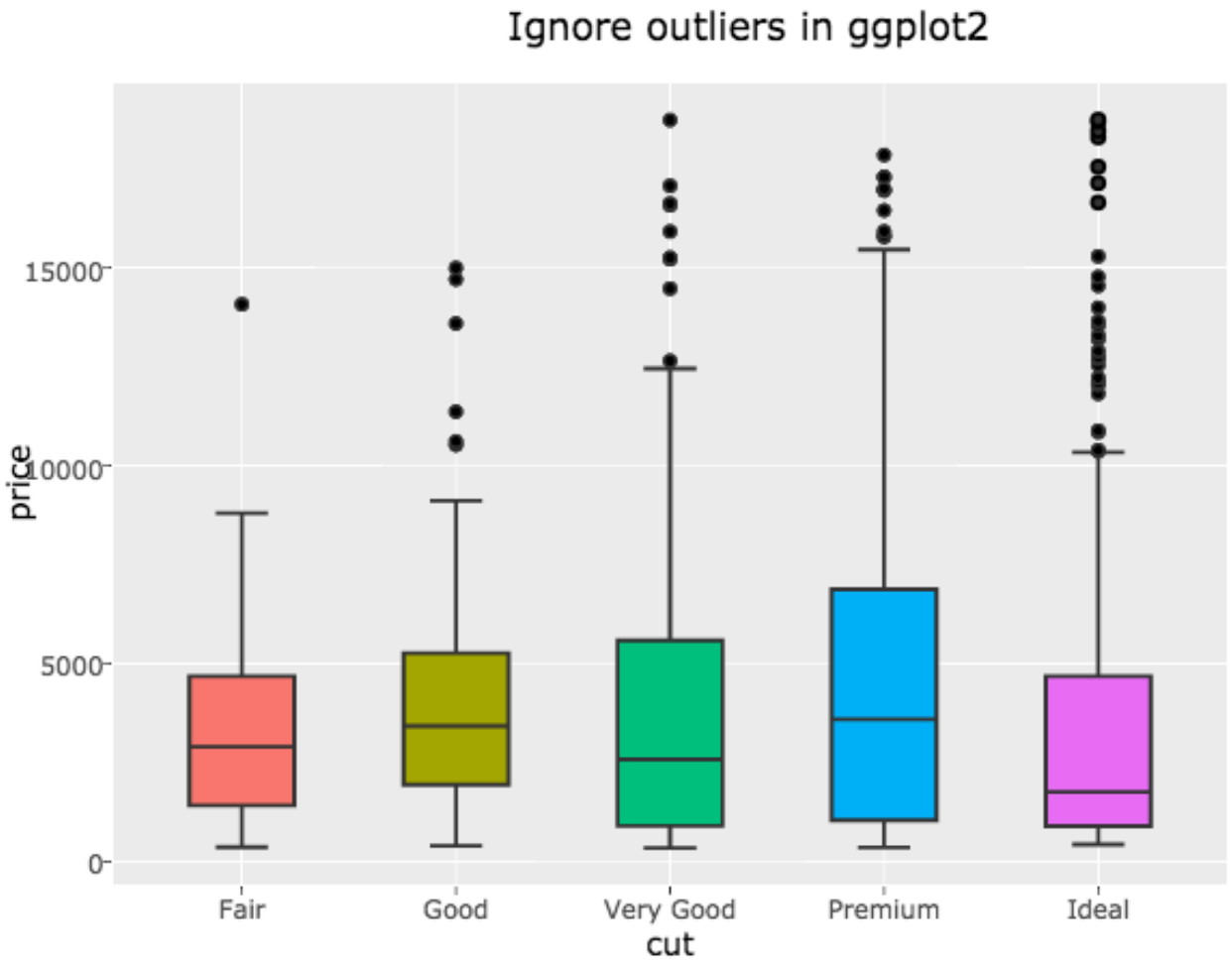
BOXPLOT:



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de posição e dispersão

BOXPLOT e OUTLIERS:



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

Qual a finalidade das medidas de dispersão?

Determinar um valor que resuma como é a variabilidade do conjunto de dados, ou seja, como eles estão distribuídos

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

AMPLITUDE:

Amplitude = (valor máx) – (valor min)

Quanto maior a amplitude amostral, maior tende a ser a dispersão de nossos dados.

Por exemplo, sejam as idades de $n = 8$ pessoas, em anos completos:

28, 20, 29, 27, 23, 27, 24 e 24.

A pessoa mais velha tem 29 anos, e a mais nova, 20 anos. A amplitude amostral é, portanto,

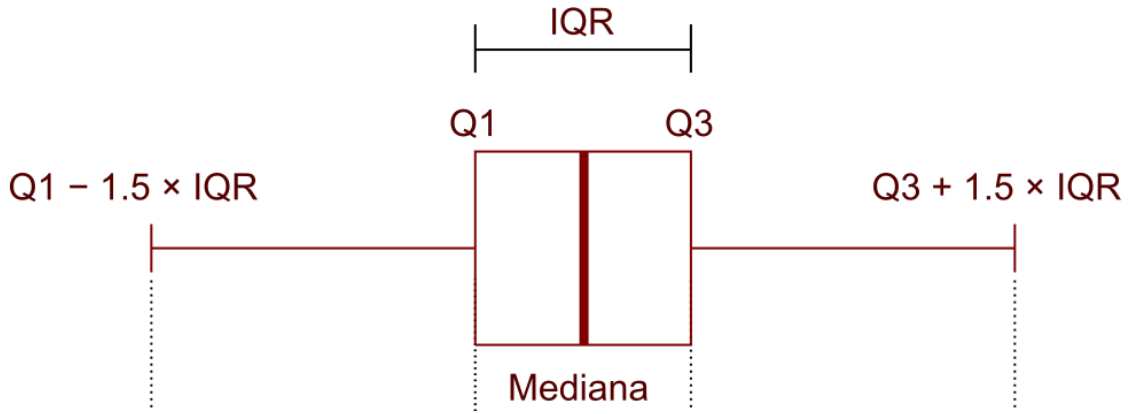
$$29 - 20 = 9 \text{ anos.}$$

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

DESVIO INTERQUARTÍLICO:

IQR = (quartil 3) – (quartil 1)



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

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

VARIÂNCIA $(\sigma^2$ ou $s^2)$:

Outro ponto importante a ser considerado diz respeito à interpretação do valor encontrado. Observe que a unidade de medida da variância é a mesma unidade de medida da variável original, mas elevada ao quadrado.

Mas o que é um ano ao quadrado, por exemplo? Como não conseguimos interpretar objetivamente essa nova unidade, é conveniente eliminarmos esse "ao quadrado". Ao fazermos isso, definimos o desvio padrão.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

DESVIO-PADRÃO (σ ou s):

O desvio-padrão é a raiz quadrada da variância:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x_i})^2}{n-1}}$$

Quanto maior o desvio-padrão maior é a dispersão dos dados

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

COEFICIENTE DE VARIAÇÃO (CV):

É o percentual de dispersão em relação à media:

$$CV = \frac{desvio\text{-}padrão}{média} \cdot 100 \text{ (em \%)}$$

Vantagens:

>> Elimina o efeito da magnitude dos dados;

>> Determina a variação percentual em relação a média;

>> É apropriado para comparar a dispersão dos dados de duas ou mais variáveis.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

Analise as estatísticas descritivas dessas duas variáveis referentes a um grupo de crianças. Quais conclusões podemos deduzir?

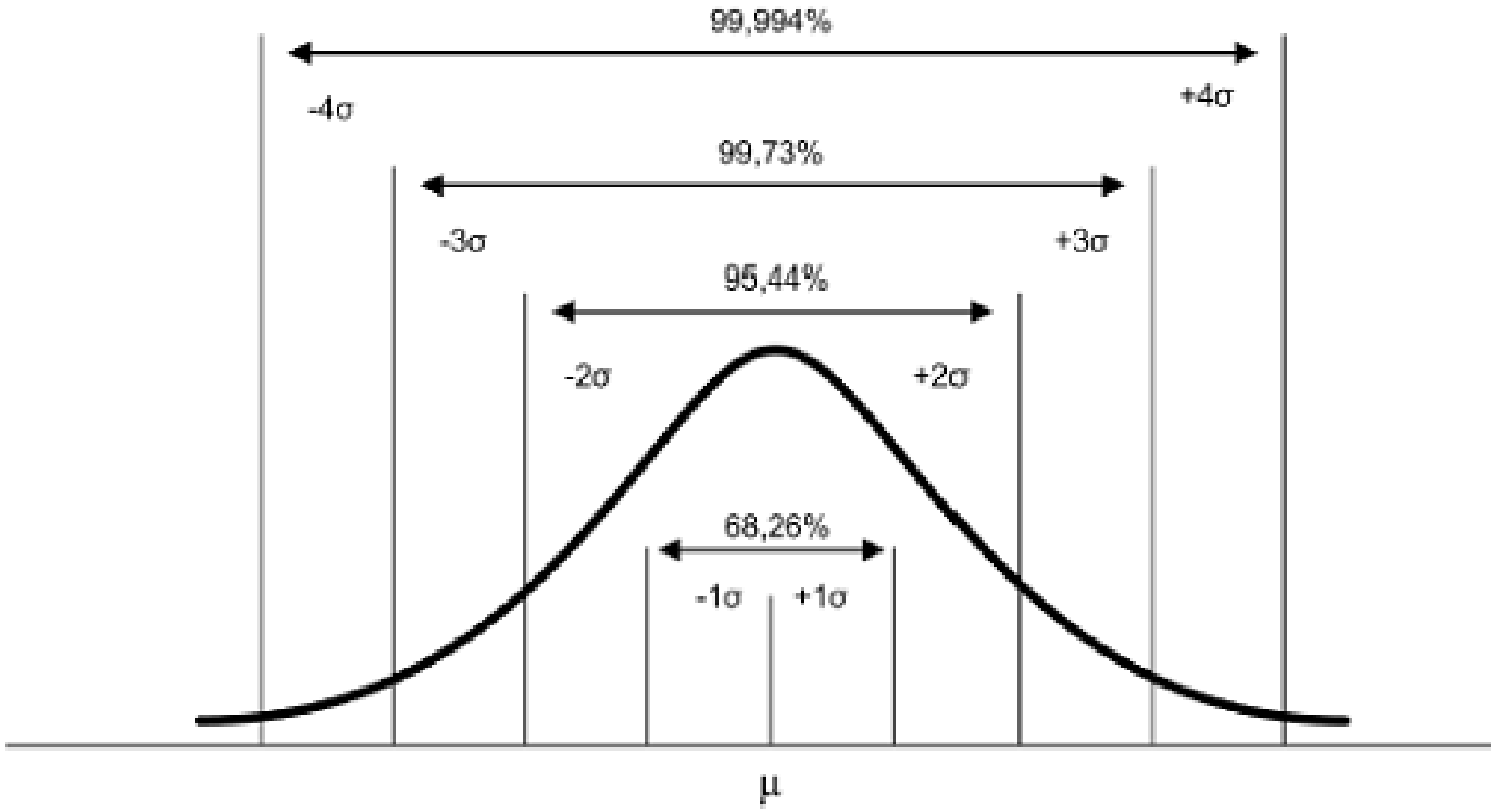
VARIÁVEIS	Média	Desvio padrão	Coeficiente de variação
Altura	1,143m	0,063m	5,5%
Peso	50kg	6kg	12%

>> Conclusão:

Em relação as médias, verificamos que nesse grupo de alunos, o peso está mais distribuído em relação a altura. O dispersão do dados de peso é quase o dobro da dispersão dos dados de altura.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

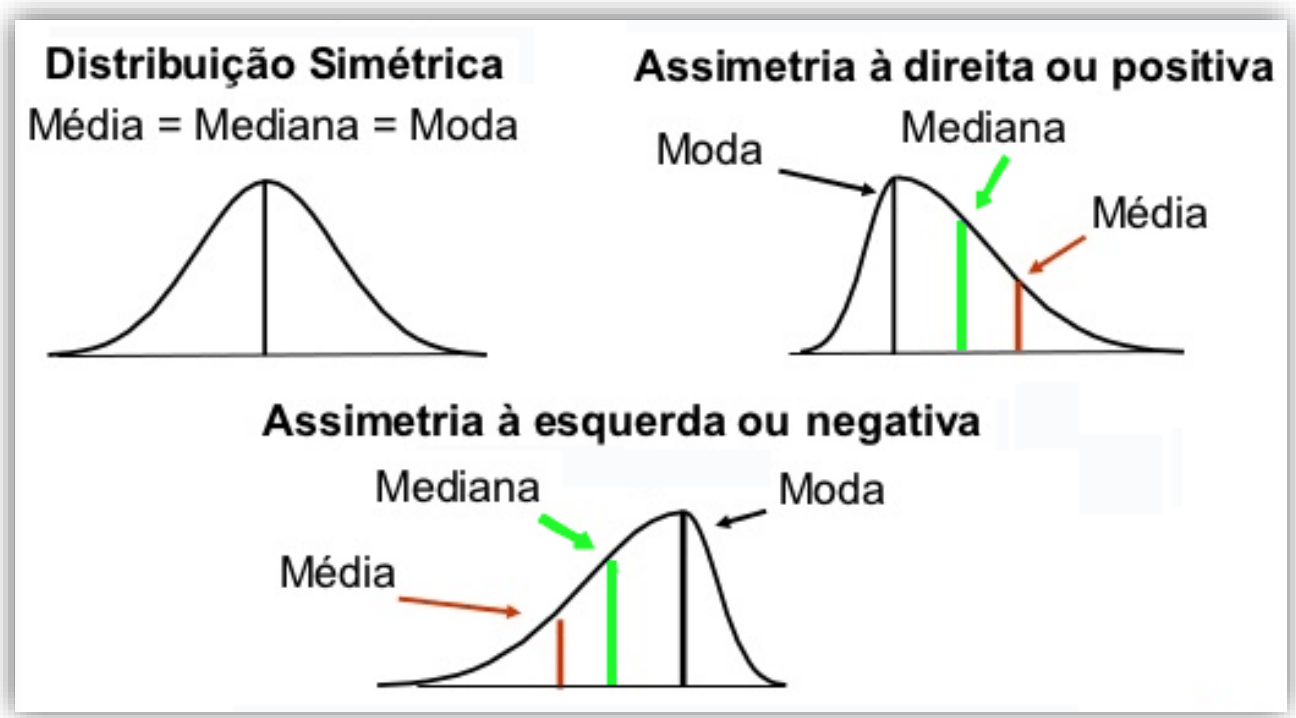
Medidas de dispersão



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

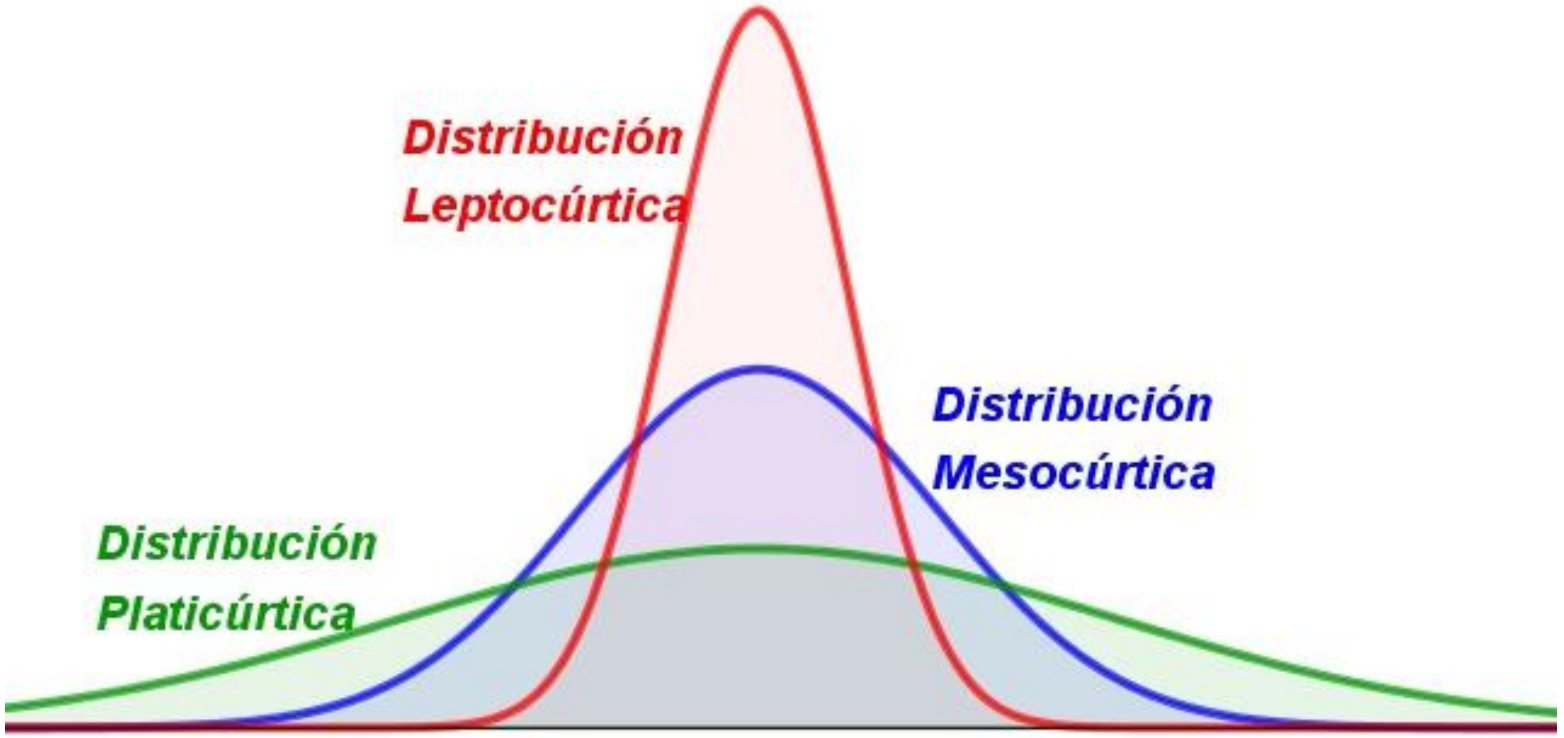
- **Assimetria:** representa a concentração dos valores em um dos extremos da distribuição



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Medidas de dispersão

- **Curtose:** representa o grau de achatamento da distribuição



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Escopo da disciplina:

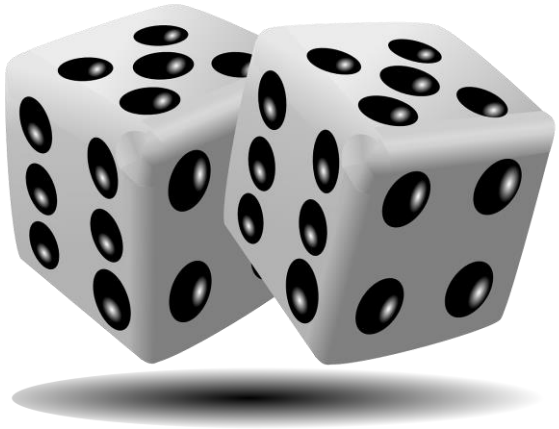
Por que e como pensar analiticamente?

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Distribuição de probabilidades

Inicialmente, o que é probabilidade?

Chance de algo acontecer!



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Distribuição de probabilidades

O que é distribuição de probabilidades?

É um modelo matemático que, através de um banco de dados (em nosso caso), **relaciona um certo valor de uma variável com a sua probabilidade de ocorrência**.

A distribuição de probabilidades pode ser representada por um **histograma de probabilidades**.

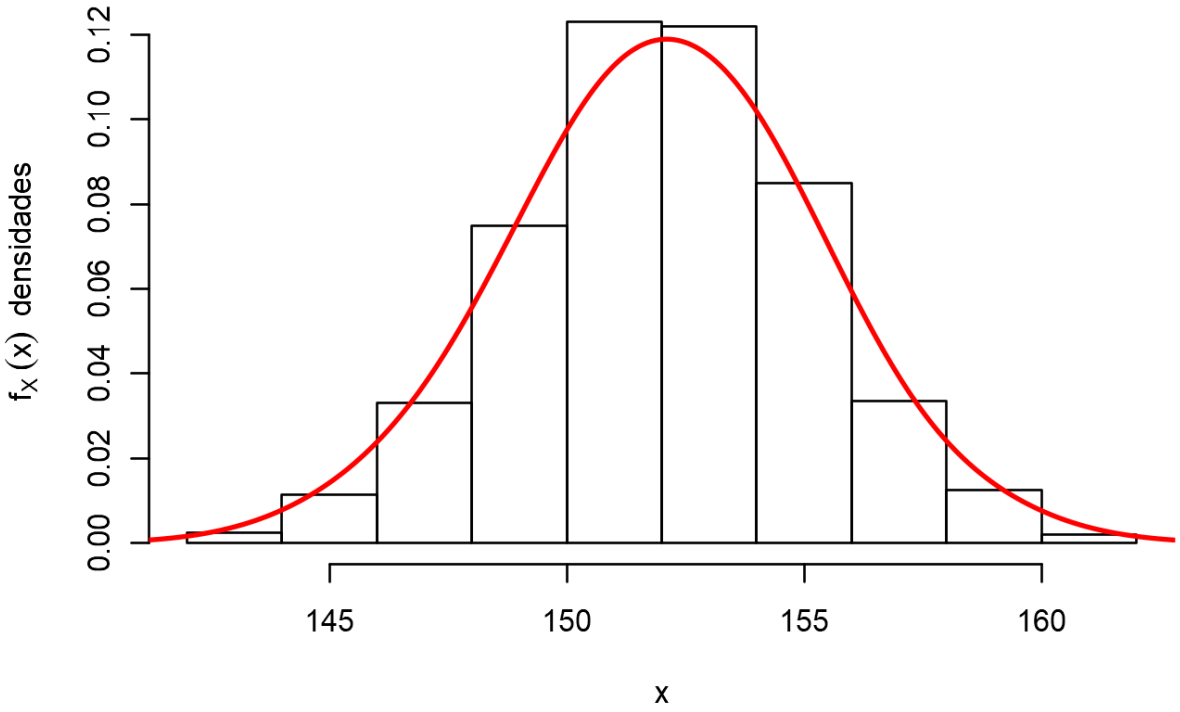
Prof. Jonathan Gil Müller <jgmuller@furb.br>

Distribuição de probabilidades

O que é distribuição de probabilidades?

É um modelo matemático que, através de um banco de dados (em nosso caso), **relaciona um certo valor de uma variável com a sua probabilidade de ocorrência**.

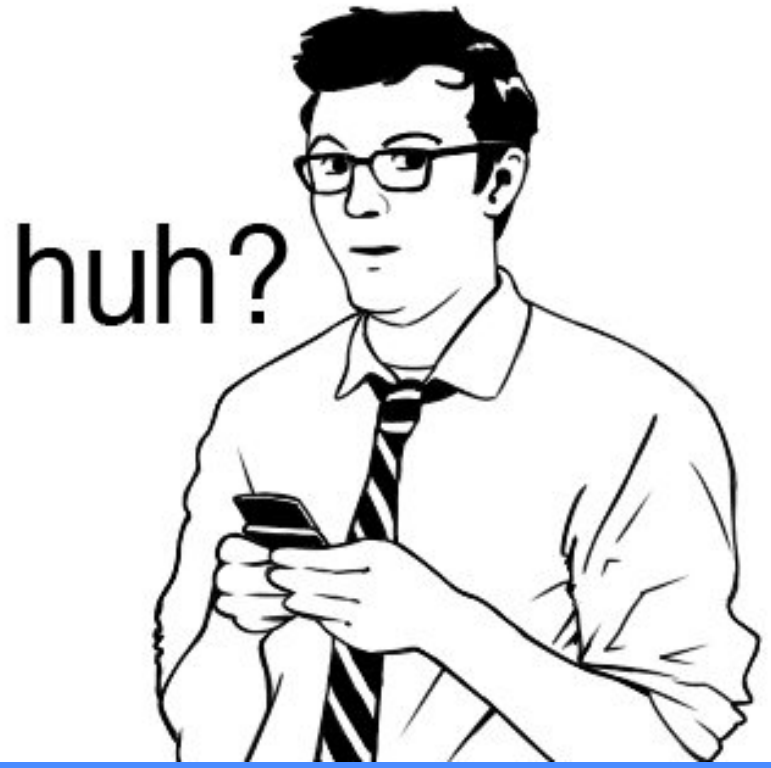
A distribuição de probabilidades pode ser representada por um **histograma de probabilidades**.



Prof. Jonathan Gil Müller <jgmuller@furb.br>

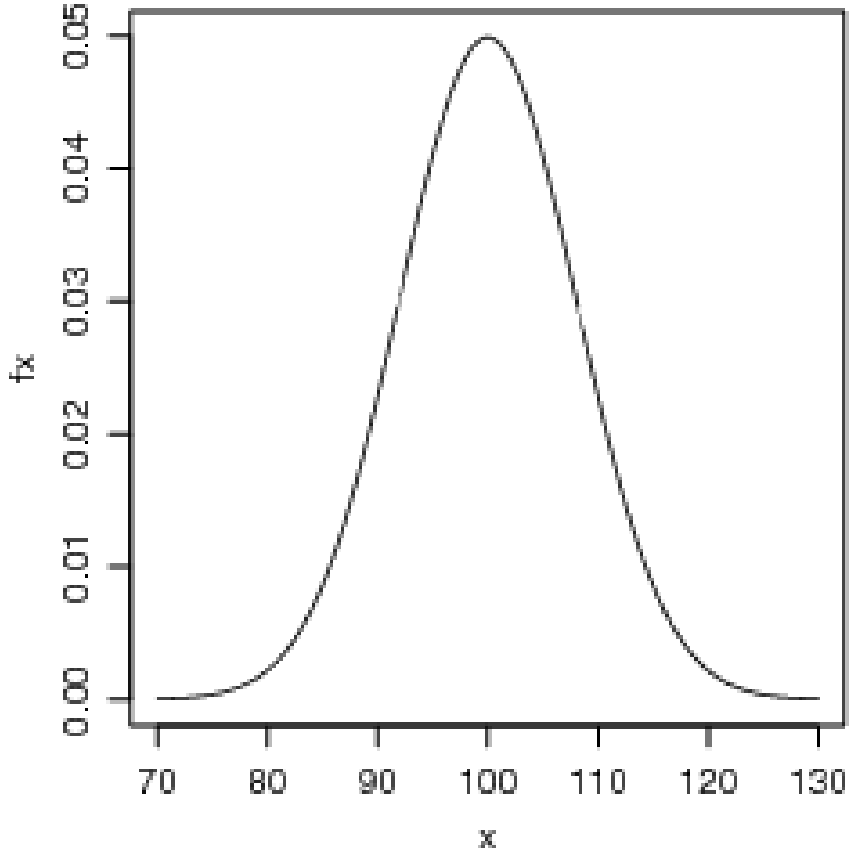
Distribuição de probabilidades

- Conhecendo como os dados de uma variável se distribuem, podemos **associar uma probabilidade para cada valor de uma variável aleatória** (aleatória pois só saberemos o seu valor depois do evento acontecer).
- Nesses caso, o que podemos fazer antes do evento acontecer é **estimar uma probabilidade com base na distribuição das informações** que já temos.
- Isso pode ajudar na **tomada de decisão em situações aleatórias.**

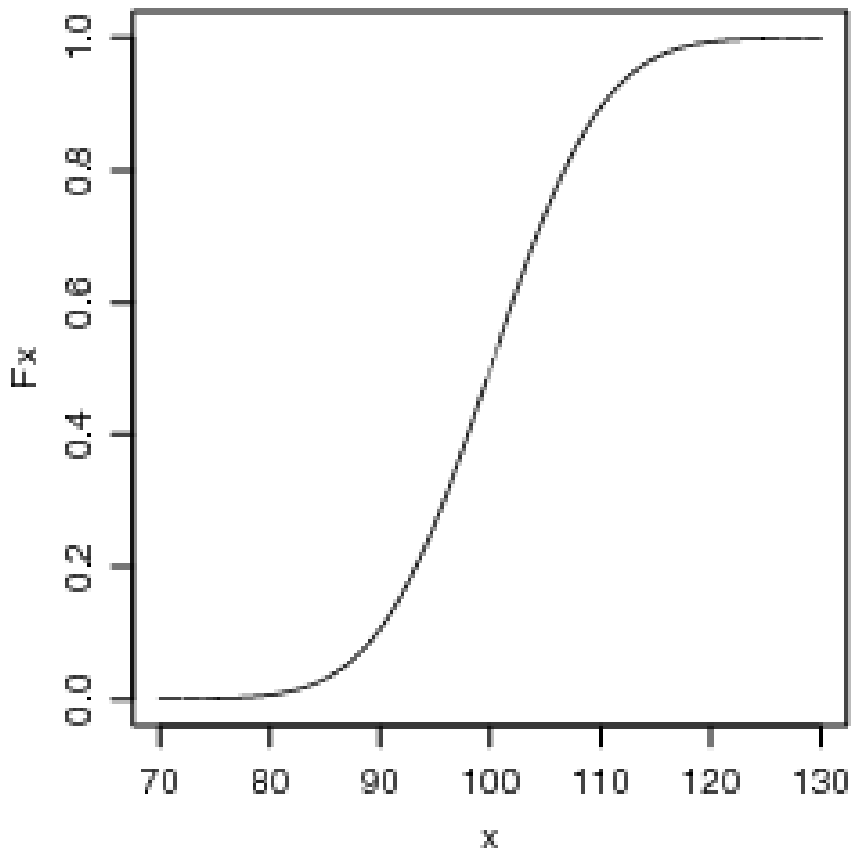


Prof. Jonathan Gil Müller <jgmuller@furb.br>

Distribuição de probabilidades



Função de densidade de probabilidade



Função cumulativa de probabilidades

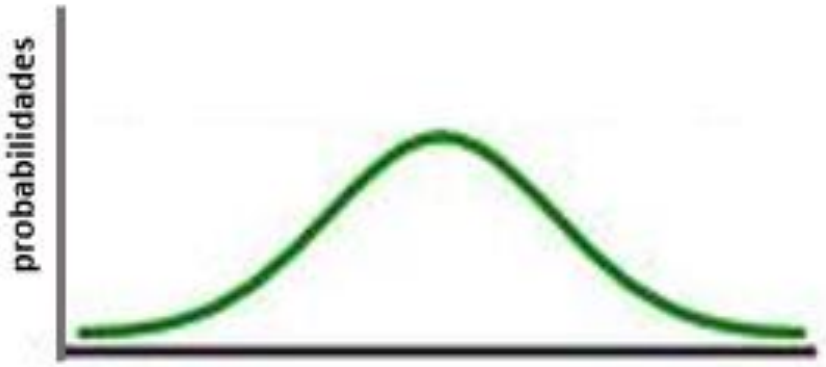
Prof. Jonathan Gil Müller <jgmuller@furb.br>

Distribuição de probabilidades

Distribuições discretas: quando a variável medida só assume valores inteiros. Nesse caso são especificadas as probabilidades para que a variável assuma um valor específico.

Distribuições contínuas: quando a variável medida é expressa em escala contínua. Nesse caso as probabilidade são especificadas em termos de intervalos.

Distribuição de probabilidades discretas



Distribuição de probabilidades contínua

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Escopo da disciplina:

Por que e como pensar analiticamente?

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Inferência estatística

Inferir significa concluir sobre ou deduzir pelo raciocínio. Este é o terceiro ramo da Estatística, em que envolve a formulação de certos julgamentos (conclusões/deduções) sobre um todo (população) após examinar apenas uma parte dele (amostra). Uma alternativa para realizar Inferência Estatística é por meio de teste de hipóteses, mas como os demais tipos de inferência, está sujeita a erro. Ela tem como base, a Teoria das Probabilidades.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Inferência estatística

A estimação de parâmetros (medidas de tendência central e dispersão) e o teste de hipótese são duas áreas importantes da inferência estatística para a tomada de decisão.

A estimação visa determinar parâmetros para a população a partir da análise de dados amostrais.

Os testes de hipótese referem-se a um processo capaz de afirmar, a partir de dados amostrais, se a hipótese de uma pesquisa é correta ou não.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

O que é uma hipótese?

Hipótese é uma afirmação que está condicionada a uma determinada população e pressupõe uma resposta à questão de pesquisa investigada.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Tipos de hipótese

(a) Hipótese Nula (H_0): quando se admite não haver diferença entre a informação fornecida pela realidade e a afirmação da hipótese.

(b) Hipótese Alternativa (H_1): quando se admite haver diferença entre a informação fornecida pela realidade e a afirmação da hipótese. É a hipótese do pesquisador.

Assim, o processo de teste consiste em aceitar ou rejeitar a hipótese nula (H_0) com base na diferença entre o valor hipotético e seu valor estimado.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Tipos de erro em testes de hipótese

O critério adotado para rejeitar ou aceitar a hipótese nula, com base em evidência amostral, não garante uma conclusão correta. Na verdade, tal decisão sempre envolve erro. Dois tipos de erros podem ocorrer:

(a) Erro tipo I: rejeição da hipótese nula (H_0), quando esta for verdadeira,
(b) Erro tipo II: não-rejeição (aceitação) da hipótese nula (H_0), quando esta for falsa.

		A verdade sobre H_0	
		H_0 é verdadeira	H_0 é falsa
Decisão	Rejeitamos H_0	Erro tipo I	Não há erro
	Não rejeitamos H_0	Não há erro	Erro tipo II

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Tipos de erro em testes de hipótese

Considere um **exemplo** em que um biomédico sintetiza uma nova droga para o tratamento de uma doença específica. Há uma droga conhecida, que chamaremos aqui de "droga 1", que é rotineiramente utilizada no tratamento da doença em questão. Vamos chamar a nova droga de "droga 2". O biomédico decide conduzir um estudo em que uma amostra de portadores da doença será aleatoriamente dividida em dois grupos. Os indivíduos do primeiro grupo receberão a droga 1, e os indivíduos do segundo grupo receberão a droga 2. O biomédico propõe que, se a droga 2 produzir melhores resultados, deverá substituir a droga 1.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Tipos de erro em testes de hipótese

Sejam θ_1 e θ_2 as proporções de respostas às drogas 1 e 2, respectivamente, em uma população de pessoas portadoras da doença. As hipóteses nula e alternativa são, respectivamente:

$$H_0 : \theta_1 \geq \theta_2$$

$$H_A : \theta_1 < \theta_2$$

Um erro tipo I consiste em concluir que a droga 2 é superior à droga 1 sem que isso seja verdade. Um erro tipo II consiste em não descartar a possibilidade de a droga 1 ser igual ou superior à droga 2, ao passo que a droga 2 é realmente superior.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Tipos de erro em testes de hipótese

Nesse exemplo, **qual erro é o mais grave**? O erro tipo I ou o erro tipo II? Para julgarmos qual erro é o mais grave, vamos entender as consequências de cada um.

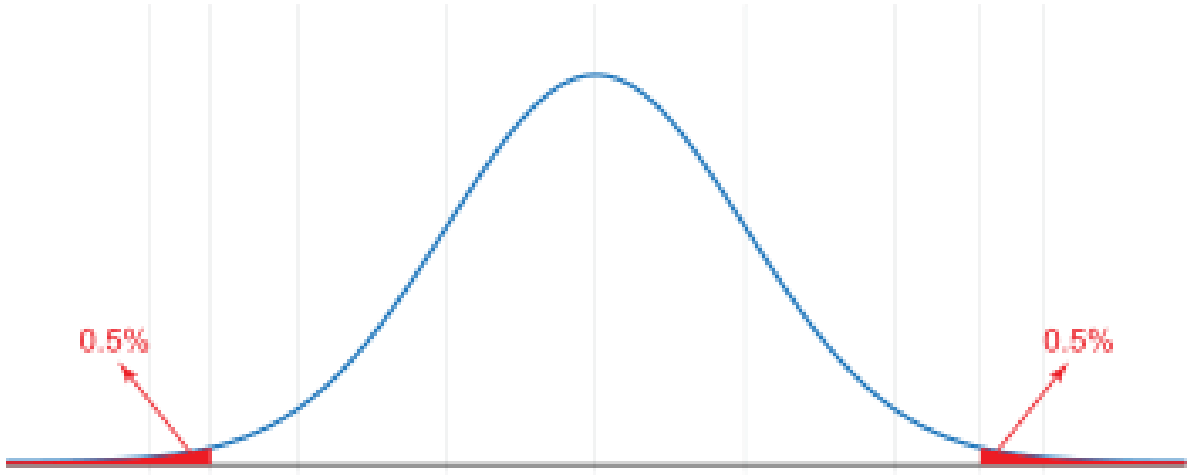
Se cometermos um erro tipo I, vamos acreditar que a droga 2 é superior, e substituir uma droga já conhecida por outra é bastante prejudicial ao tratamento dos pacientes. Se cometermos um erro tipo II, não teremos argumentos para substituir a droga 1 pela 2, e tudo continuará como estava antes. Como prejuízo, perderemos a chance de ter evidências de que há uma droga melhor que aquela rotineiramente utilizada.

Assim, entendemos que **nesse exemplo o erro tipo I é mais grave**. Substituir um conhecimento tradicional por um alternativo, sendo o alternativo falso, é mais prejudicial que deixar de substituir um conhecimento tradicional por um alternativo, sendo o alternativo verdadeiro. Por isso, em muitas ocasiões, o erro tipo I é mais grave que o erro tipo II.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Nível de significância

Na construção de um teste de hipóteses, procuramos controlar o erro tipo I fixando o **valor de α (probabilidade de cometer o erro tipo I).** Entendemos, então, que o **nível de significância** não é um valor que calculamos a partir da amostra, mas é escolhido quando nosso estudo está sendo planejado e **representa um limite (para aceitar a hipótese nula como como verdadeira) da probabilidade de obter uma amostra com a estatística da hipótese nula.**

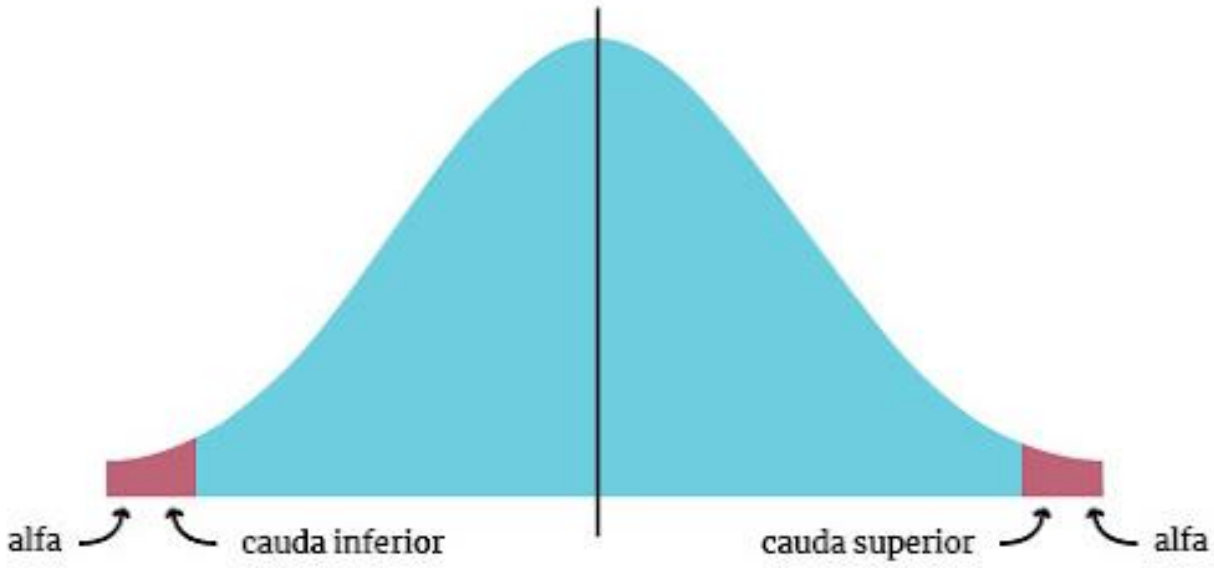


Prof. Jonathan Gil Müller <jgmuller@furb.br>

Valor-p (*p-valor*)

O valor-p é uma quantificação da probabilidade de se errar ao rejeitar H_0 e a mesma decorre da distribuição estatística adotada. Podemos também dizer, que o **p-valor representa a probabilidade de assumirmos que a H_0 é verdadeira.**

Se o valor-p é menor que o nível de significância, conclui-se que o correto é rejeitar a hipótese de nulidade e vice-versa.

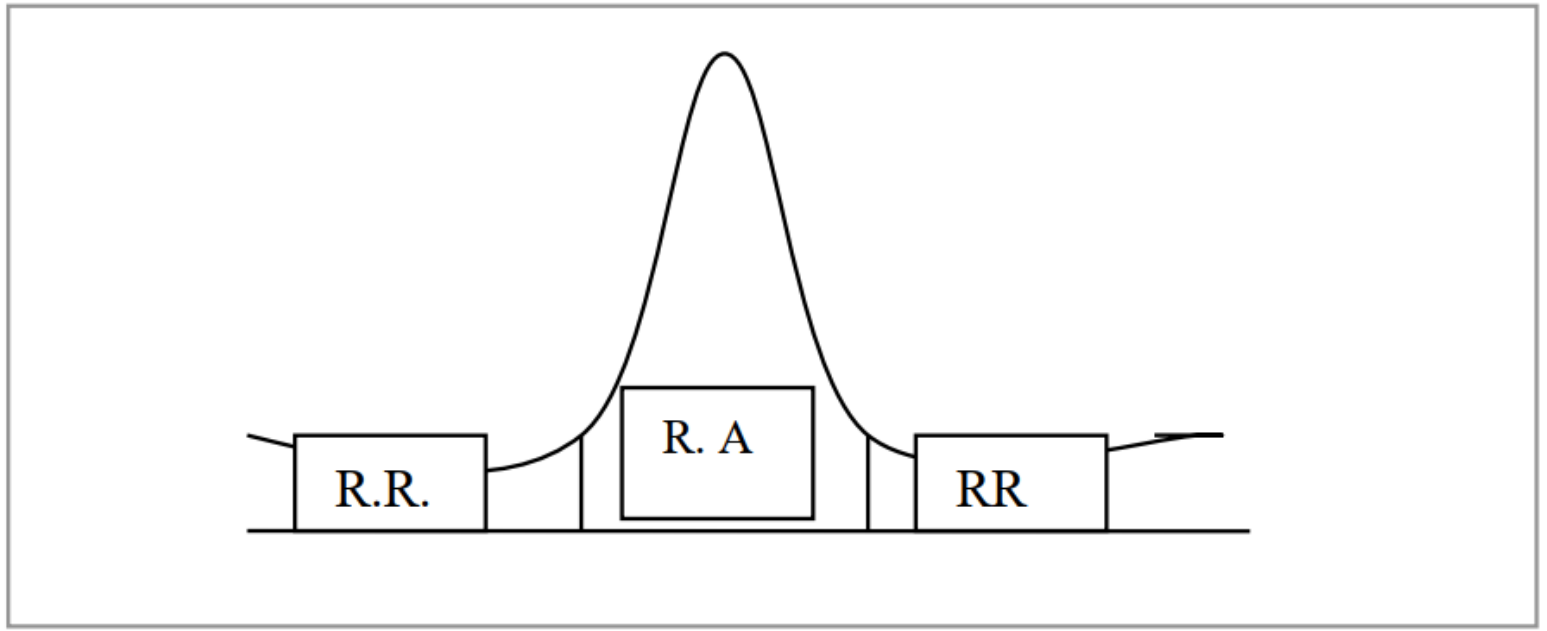


Regiões críticas

Teste bicaudal

H0: μ = X (ausência do efeito)

H1: μ ≠ X (presença do efeito positivo ou negativo)



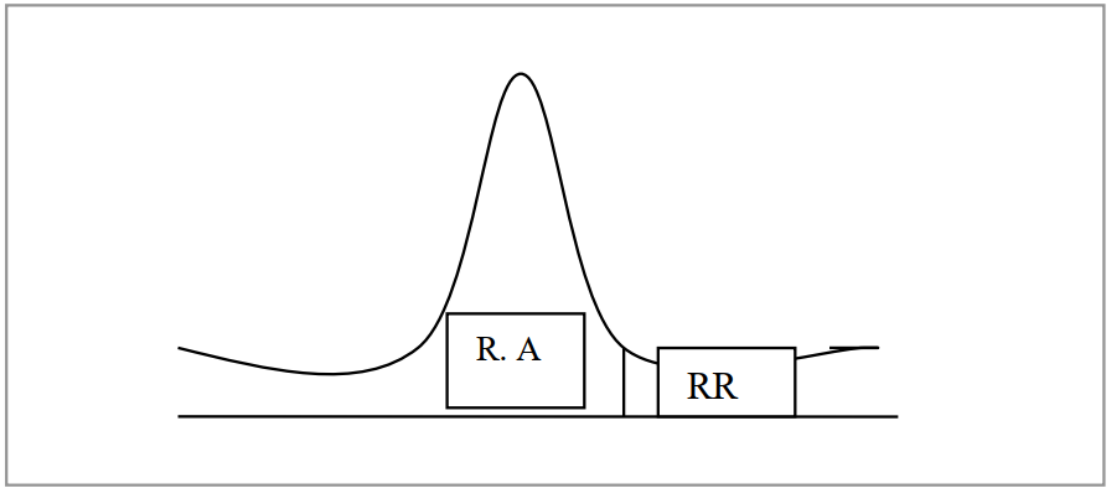
Prof. Jonathan Gil Müller <jgmuller@furb.br>

Regiões críticas

Teste monocaudal

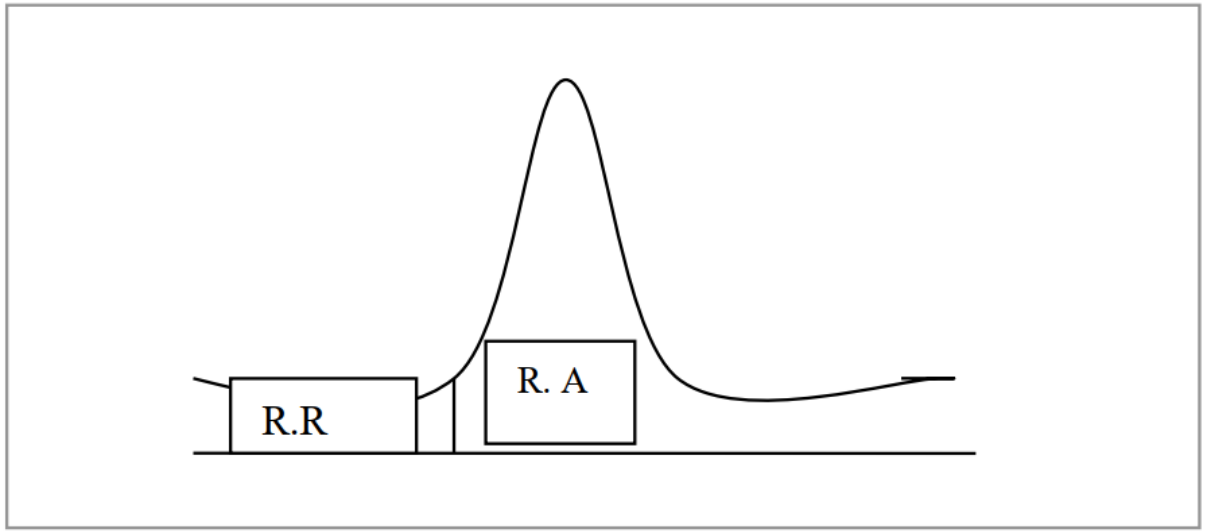
H0: μ = X (ausência do efeito)

H1: μ > X (presença do efeito positivo)



H0: μ = X (ausência do efeito)

H1: μ < X (presença do efeito negativo)



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Teste de hipóteses

Para realizar um teste de hipóteses e divulgar as conclusões é necessário seguir um procedimento aceito pela comunidade científica.

Neste procedimento, o pesquisador deve deixar claro qual a hipótese que ele deseja testar. Para isto ele precisa escrever em termos estatísticos a sua hipóteses cientifica (H_A).

A hipótese científica do pesquisador, nada mais é o que o levou a realizar a sua investigação.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Teste t de Student

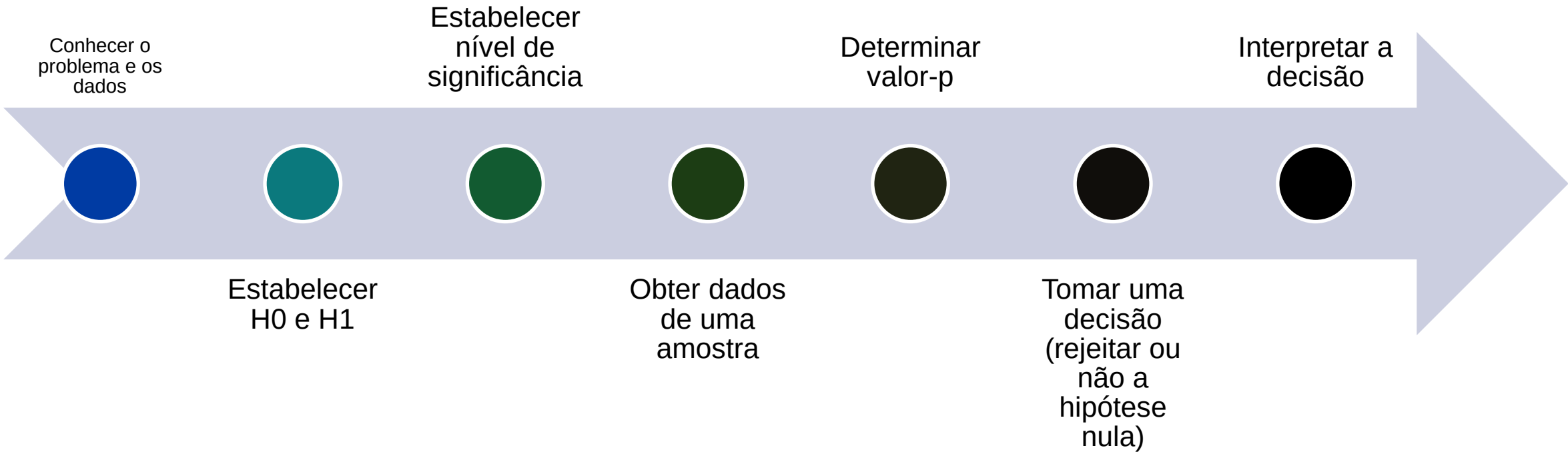
O teste t para uma média populacional (pode ser aplicado para duas ou mais médias populacionais) é um teste estatístico de hipótese usado para verificar se a média de uma característica de uma população assume o valor especificado. Esse teste pode ser aplicado quando:

- conhecemos o s (desvio-padrão);
- quando o tamanho da amostra for menor que 30 ($n<30$).

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Teste t de Student

Orientações para a execução de um teste de hipótese



Prof. Jonathan Gil Müller <jgmuller@furb.br>

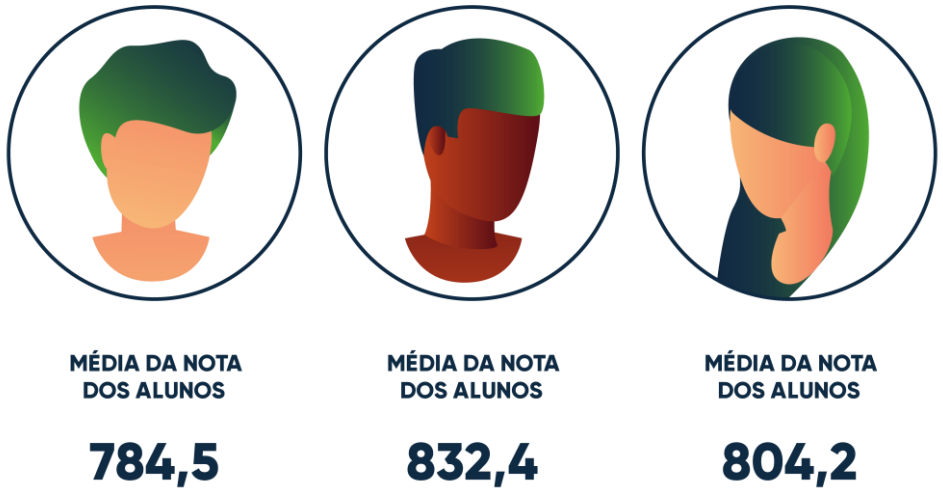
Teste Z

O teste Z é um **teste estatístico para uma média populacional**. O teste Z pode ser usado quando a população for normal, σ (desvio-padrão populacional) for conhecido e o **tamanho da amostra for maior ou igual a 30**.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Análise de Variância (ANOVA)

Trata-se de uma técnica estatística utilizada para avaliar médias populacionais. Esta análise busca **identificar a existência de uma diferença significativa entre as médias e esses fatores exercem influência em alguma variável dependente**.



Prof. Jonathan Gil Müller <jgmuller@furb.br>

Escopo da disciplina:

Por que e como pensar analiticamente?

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Como pensar analiticamente?

Existem maneiras diferentes que variam de acordo com o tipo dos dados analisados e com tipo da análise a ser feita.

A **compreensão de conceitos matemáticos e estatísticos alinhados ao domínio de ferramentas tecnológicas** facilitam a análise de grandes volumes de dados.

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Por que pensar analiticamente

No mundo atual, temos diversos equipamentos coletando dados e informações continuadamente. A análise dessas informações possibilita o **desenvolvimento de novos conhecimentos e a solução de problemas emergentes do século XXI.**

Prof. Jonathan Gil Müller <jgmuller@furb.br>

Muito obrigado!