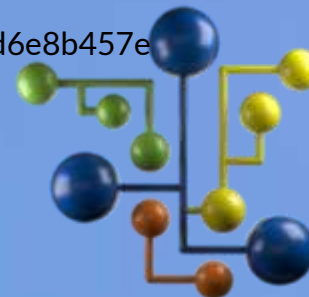


Data Science  
Academy

Data Science Academy [adeilsonjunior@gmail.com](mailto:adeilsonjunior@gmail.com) 5aabf79e5e4cdefd6e8b457e



# Big Data Analytics com R e Microsoft Azure Machine Learning



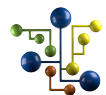
Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Big Data Analytics com R e Microsoft Azure Machine Learning

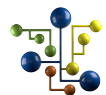
## Introdução à Análise Estatística de Dados

Seja Bem-Vindo(a)!



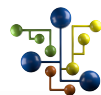
# Introdução à Análise Estatística de Dados

Iniciamos agora a segunda parte do curso em que vamos estudar Análise Estatística e Machine Learning.



# Introdução à Análise Estatística de Dados

Estatística é uma área muito ampla e este capítulo vai oferecer a você uma introdução aos principais conceitos usados em Data Science.



# Introdução à Análise Estatística de Dados

E muitos desses conceitos serão explorados nos capítulos seguintes, quando estudarmos Machine Learning.



# Introdução à Análise Estatística de Dados

Mas lembre-se: Data Science envolve muitas outras áreas além da Estatística, como Matemática, Ciência da Computação (Programação, Armazenamento e Processamento de Dados) e conhecimento sobre áreas de negócio.

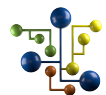


# Introdução à Análise Estatística de Dados

Introdução à  
Análise Estatística  
de Dados  
Parte 1

Introdução à  
Análise Estatística  
de Dados  
Parte 2

Introdução à  
Análise Estatística  
de Dados  
Parte 3



# Introdução à Análise Estatística de Dados

Na seção de links úteis ao final do capítulo você vai encontrar algumas recomendações de cursos complementares que poderão ajudar você a desenvolver suas habilidades em Análise Estatística.

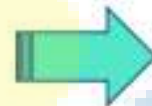
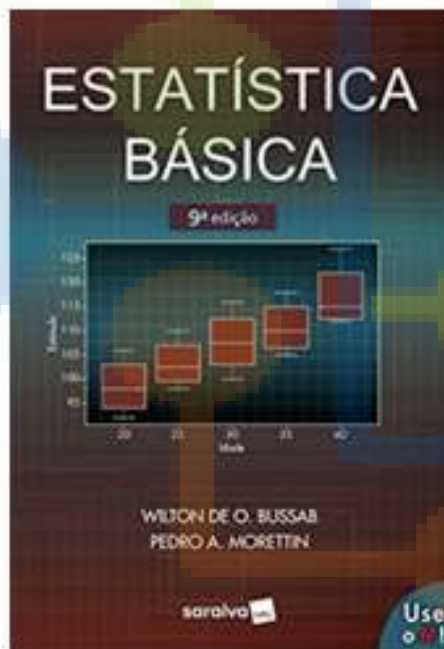




Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Introdução à Análise Estatística de Dados





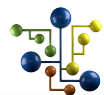
Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Big Data Analytics com R e Microsoft Azure Machine Learning

## O Papel da Estatística em Ciência de Dados

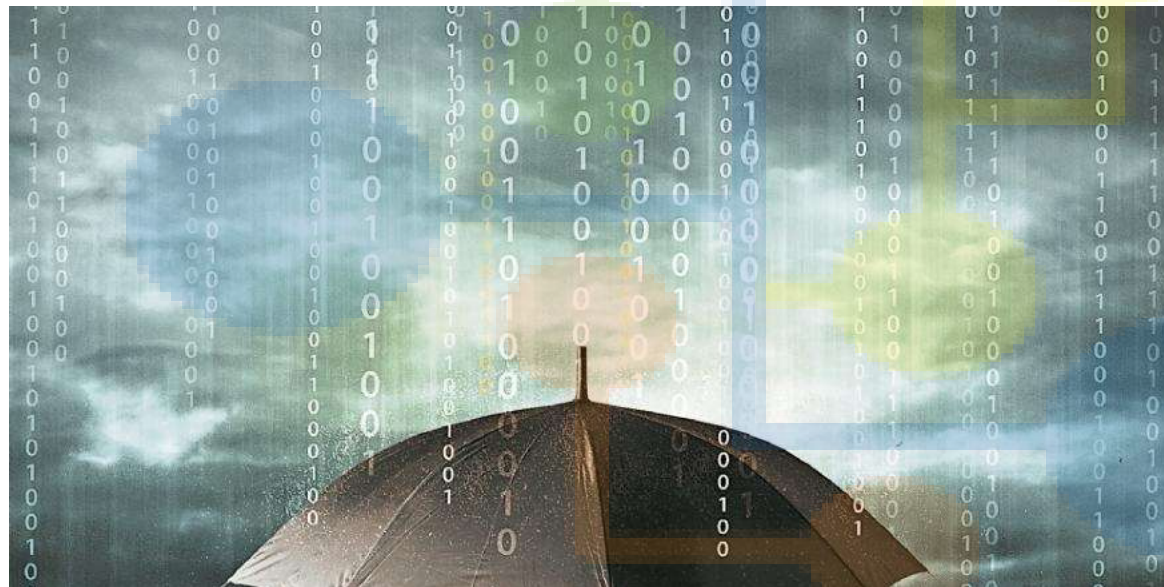
### Seja Bem-Vindo(a)!

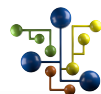


Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# O Papel da Estatística em Ciência de Dados

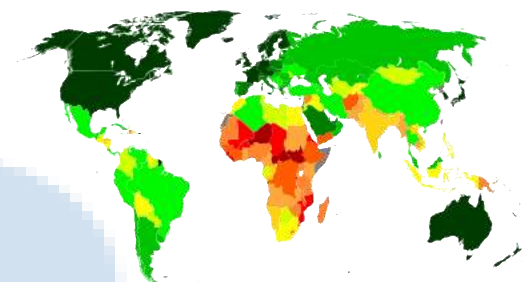


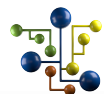


Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

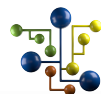
# O Papel da Estatística em Ciência de Dados





# O Papel da Estatística em Ciência de Dados

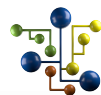
Tudo isso é o que chamamos de dados e, sozinhos, dados são apenas um apanhado de ruído e confusão.



# O Papel da Estatística em Ciência de Dados

Para dar-lhes sentido, interpretações e significados, necessita-se de um ramo poderoso da ciência: Estatística, que é fundamental no processo de descoberta científica ao fornecer modelos capazes de aprimorar pesquisas e nortear tomadas de decisão.



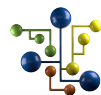


# O Papel da Estatística em Ciência de Dados

Ao observar um fenômeno sucessivamente é possível notar que, muito raramente, os resultados encontrados serão iguais.

Isso porque, praticamente tudo está sujeito a variação.



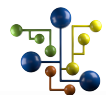


# O Papel da Estatística em Ciência de Dados

No entanto, o uso de métodos estatísticos permite que se facilite a compreensão e descrição dessa “inconstância” e que ela seja usada de forma a ajudar no processo de tomada de decisão.



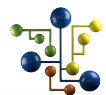




# O Papel da Estatística em Ciência de Dados

A Estatística possui diversas aplicações e cumpre os mais variados objetivos, sendo especialmente útil e, por vezes, indispensável quando se trata da Ciência de Dados.

Usamos Estatística para descrever, resumir e explorar os dados. Ao trabalhar com Machine Learning, usamos Estatística para interpretar e avaliar os resultados do modelo.



# O Papel da Estatística em Ciência de Dados

Em Ciência de Dados, a Estatística cumpre um papel importante. Mas usamos ainda Matemática para criar um modelo de Machine Learning e Ciência da Computação quando precisamos criar programas de software para análise ou armazenar e processar os dados de forma distribuída, o que é necessário quando o volume de dados é muito grande. Isso sem falar no conhecimento das áreas de negócio.

Data Science envolve muitas áreas e a Estatística é uma delas.



Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Big Data Analytics com R e Microsoft Azure Machine Learning

As 3 Grandes Áreas da Estatística

Seja Bem-Vindo(a)!



# As 3 Grandes Áreas da Estatística

**As três  
grandes áreas  
da Estatística**



Probabilidade



Estatística Descritiva



Estatística Inferencial



# As 3 Grandes Áreas da Estatística



**Probabilidade** – estudo da aleatoriedade e da incerteza.

Utiliza métodos de quantificação das chances associadas aos diversos resultados.



# As 3 Grandes Áreas da Estatística



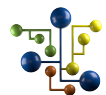
**Estatística Descritiva** – utiliza métodos para coleta, organização, apresentação, análise e síntese de dados obtidos em uma população ou amostra.



# As 3 Grandes Áreas da Estatística



**Estatística Inferencial** é o processo de estimar informações sobre uma população a partir dos resultados observados em uma amostra.



Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# As 3 Grandes Áreas da Estatística

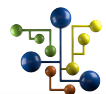






# As 3 Grandes Áreas da Estatística

**Estatística é a ciência, parte da Matemática Aplicada, que fornece métodos para coletar, descrever, analisar, apresentar e interpretar dados, para a utilização dos mesmos na tomada de decisões.**



# Big Data Analytics com R e Microsoft Azure Machine Learning

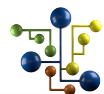
População e Amostra

Seja Bem-Vindo(a)!



# População e Amostra

Sempre que você se deparar com um novo conjunto de dados, a primeira pergunta que deve ser feita é: qual é minha população e qual é minha amostra?



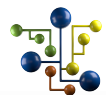
# População e Amostra

## População

São todos os elementos distintos – indivíduos, itens ou objetos – cujas características estejam sendo estudadas.

## Amostra

É uma parte da população, sendo coletada a partir da população que está sendo estudada.



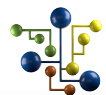
Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# População e Amostra

## Exemplo





# População e Amostra

## Exemplo





Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Big Data Analytics com R e Microsoft Azure Machine Learning

Como Garantir que a Amostra Representa  
Fielmente a População?

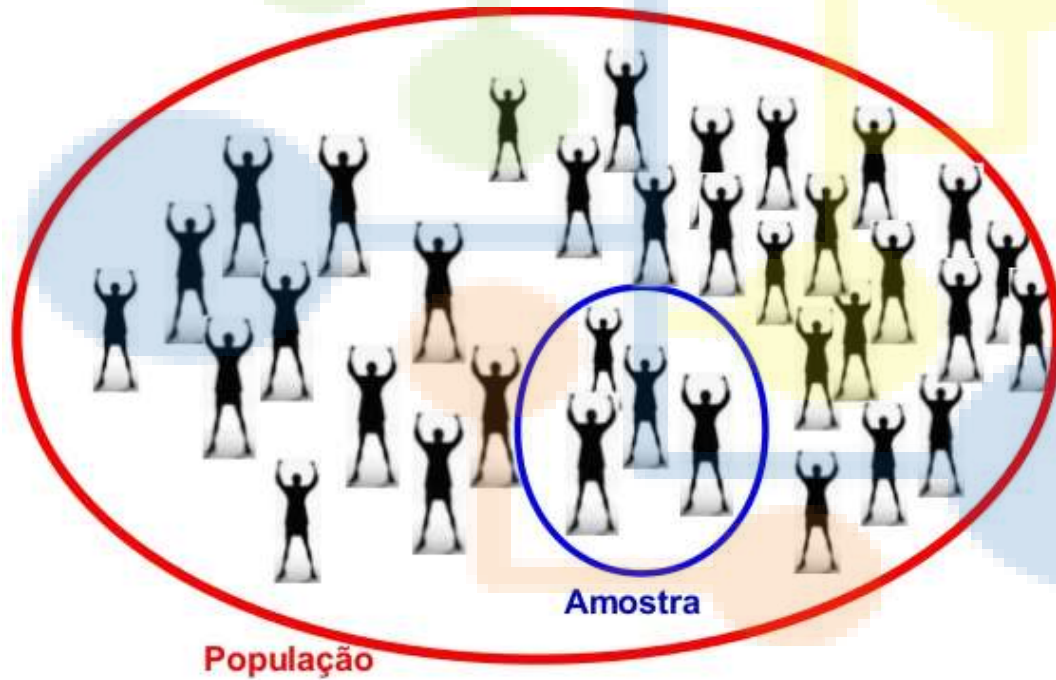
Seja Bem-Vindo(a)!



Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

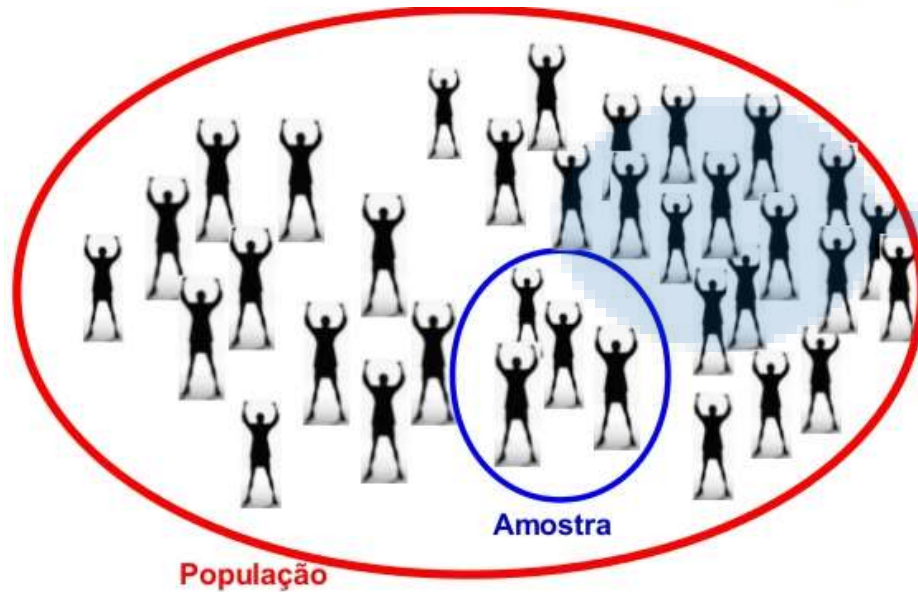
# Como Garantir que a Amostra Representa Fielmente a População?







# Como Garantir que a Amostra Representa Fielmente a População?



A melhor estratégia para criar uma amostra fiel é através de randomização ou aleatorização.



Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Como Garantir que a Amostra Representa Fielmente a População?

É Sopa Novamente!





Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Como Garantir que a Amostra Representa Fielmente a População?

É Sopa Novamente!





# Como Garantir que a Amostra Representa Fielmente a População?

Ok, entendi. Mas como eu faço com indivíduos??





# Como Garantir que a Amostra Representa Fielmente a População?

Como garanto que minha amostra está devidamente randomizada?





Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Como Garantir que a Amostra Representa Fielmente a População?

Simplesmente, você coleta sua **amostra** de forma randomizada, sem escolher exatamente quem fará parte da amostra.

Não pense que é fácil como parece. Tão importante quanto a **randomização** é o **tamanho** da **amostra** e existem diversas técnicas para a coleta de amostra de dados.



# Como Garantir que a Amostra Representa Fielmente a População?

E qual deve ser o tamanho da **amostra**?





# Como Garantir que a Amostra Representa Fielmente a População?

Suponhamos que sua empresa, na área de prestação de serviços de saúde, precisasse responder a seguinte pergunta:

As pessoas que ingerem um tipo específico de bebida alcoólica (cerveja, por exemplo) são mais suscetíveis a ter problemas e necessitarem de atendimento médico de emergência?





# Como Garantir que a Amostra Representa Fielmente a População?

Qual seria a População e qual seria a Amostra?



Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Como Garantir que a Amostra Representa Fielmente a População?

População – todos os indivíduos de um país



Amostra – pessoas que foram atendidas em um hospital específico





Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Big Data Analytics com R e Microsoft Azure Machine Learning

Parâmetros x Estatísticas

Seja Bem-Vindo(a)!



# Parâmetros x Estatísticas

As **estatísticas** se baseiam nos dados da **amostra** e não em dados populacionais. Quando se coletam dados de toda uma população, temos o chamado censo.

Se você depois resume toda a informação do censo em um número, esse número é um **parâmetro**, não uma estatística.



# Parâmetros x Estatísticas

Portanto:

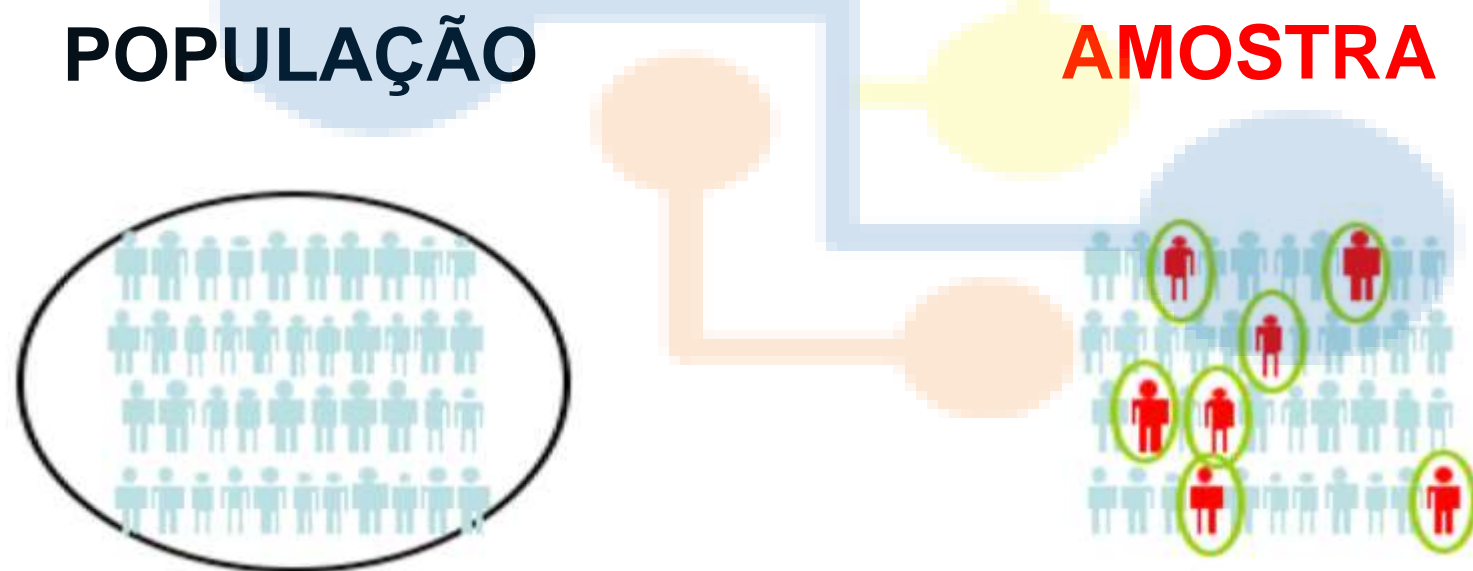
**Parâmetros** – características sobre a população. Valores calculados usando dados da população são chamados de parâmetros.

**Estatísticas** – características sobre a amostra. Valores calculados usando dados da amostra são chamados de estatísticas.



# Parâmetros x Estatísticas

**Estatística Inferencial** realiza deduções e conclusões sobre a população, com base nos resultados obtidos da análise da amostra.



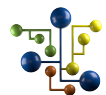


# Parâmetros x Estatísticas

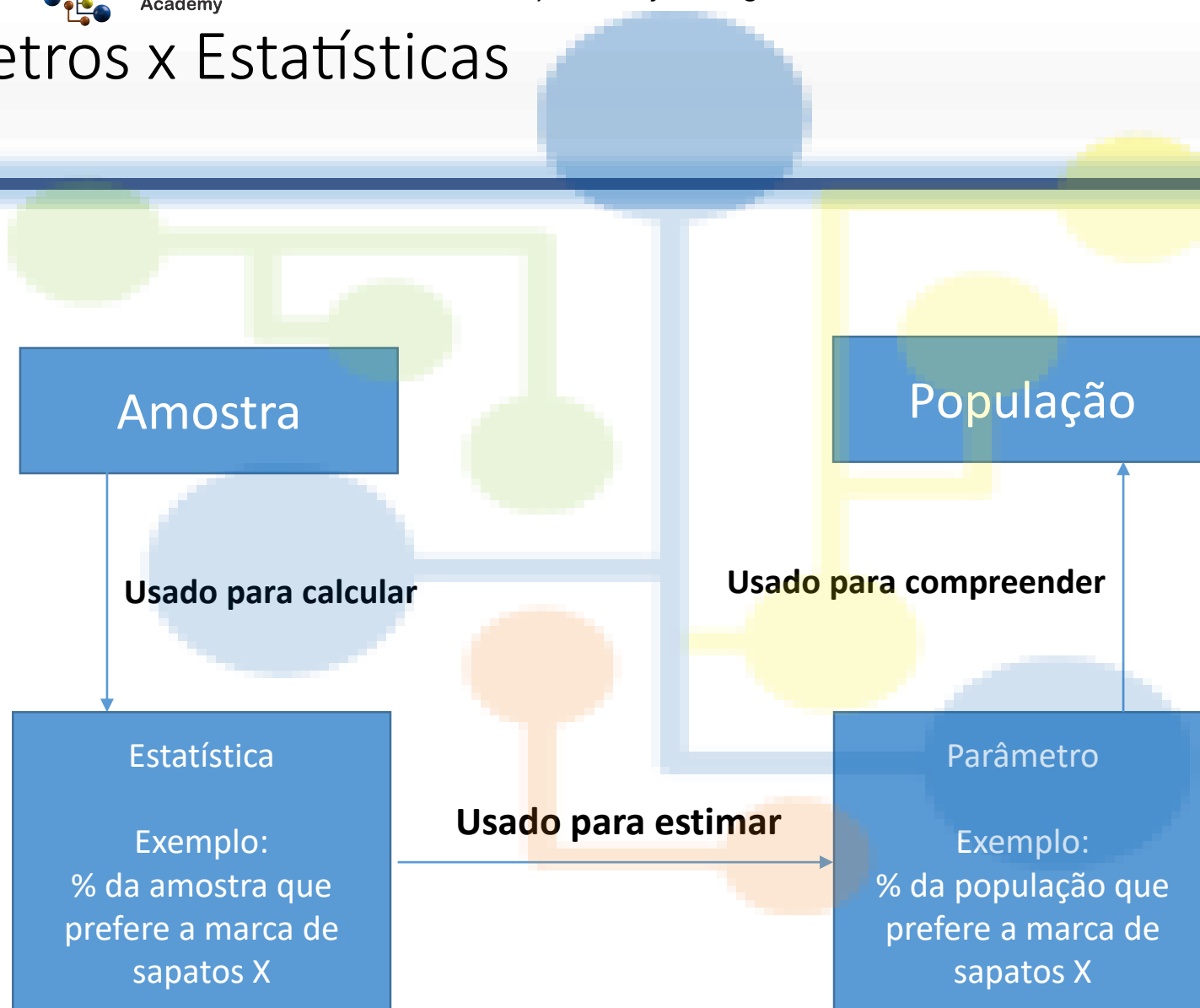
E por que não analisamos a população inteira? Por que precisamos de uma amostra?

Por diversas razões!

- Custo
- Tempo
- Necessidade



# Parâmetros x Estatísticas







Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Big Data Analytics com R e Microsoft Azure Machine Learning

Observações x Variáveis

Seja Bem-Vindo(a)!



# Observações x Variáveis

Vamos relembrar um conceito fundamental

dado



informação



# Observações x Variáveis

**Dados** – valores coletados através de observação ou medição.

**Informação** – dados que são transformados em fatos relevantes e usados para um propósito específico.



# Observações x Variáveis

Dados não fazem sentido, se não forem colocados em um contexto.



# Observações x Variáveis

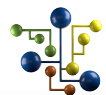
Os dados podem ser obtidos através de duas fontes principais:

## Dados Primários

- ☐ Coletados por quem faz a análise
- ☐ Confiáveis
- ☐ Possuem maior controle

## Dados Secundários

- ☐ Coletados por terceiros
- ☐ Não Confiáveis
- ☐ Não possuem muito controle



# Observações x Variáveis

Vantagens

## Dados Primários

- Confiabilidade
- Qualidade
- Controle das informações
- Acertabilidade nos resultados
- Dados atualizados

- Alto custo
- Demanda tempo maior
- Equipe grande

Desvantagens



# Observações x Variáveis

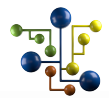
## Dados Secundários

Vantagens

- Baixo custo
- Rapidez
- Existência de diversas fontes
- Diversidade de informações para quantificação de questões

- Falta de controle
- Dados Inadequados
- Diversidade na classificação dos dados
- Dados desatualizados
- Fontes não confiáveis
- Dificuldade de reproduzir um estudo obtendo os mesmos resultados

Desvantagens



# Observações x Variáveis

Observações x Variáveis





# Observações x Variáveis

## Observação

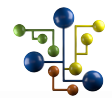
Uma observação é uma ocorrência de um item de dados específico que é gravada sobre uma unidade de dados. Também chamada de registro.



# Observações x Variáveis

## Variável

Variável é a característica de interesse que é medida em cada elemento da amostra ou população. Como o nome sugere, seus valores variam de elemento para elemento. As variáveis podem ter valores numéricos ou não numéricos.



# Observações x Variáveis

Variáveis

Observações

	Idade	Sexo	Peso	Cor dos olhos
Indivíduo 1	42	M	59	Verde
Indivíduo 2	34	M	54	Castanho
Indivíduo 3	56	F	89	Azul
Indivíduo 4	41	M	76	Castanho
Indivíduo 5	23	F	65	Castanho



Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Big Data Analytics com R e Microsoft Azure Machine Learning

Tipos de Variáveis

Seja Bem-Vindo(a)!

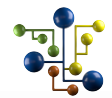


# Tipos de Variáveis

As variáveis podem ser:

**Qualitativas** – utilizam termos **descritivos** para descrever algo de interesse. Ex: cor dos olhos, estado civil, religião, sexo, grau de escolaridade, classe social, tipo sanguíneo, cor da pele, etc...





# Tipos de Variáveis

## Variáveis Qualitativas

Observações

	Idade	Sexo	Peso	Cor dos olhos
Indivíduo 1	42	M	59	Verde
Indivíduo 2	34	M	54	Castanho
Indivíduo 3	56	F	89	Azul
Indivíduo 4	41	M	76	Castanho
Indivíduo 5	23	F	65	Castanho

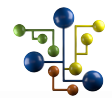


# Tipos de Variáveis

As variáveis podem ser:

**Quantitativas** – representadas por valores numéricos que podem ser **contados** ou **medidos**. Ex: número de crianças em uma sala de aula, peso do corpo humano, idade, número de filhos, etc...





# Tipos de Variáveis

## Variáveis Quantitativas



Observações



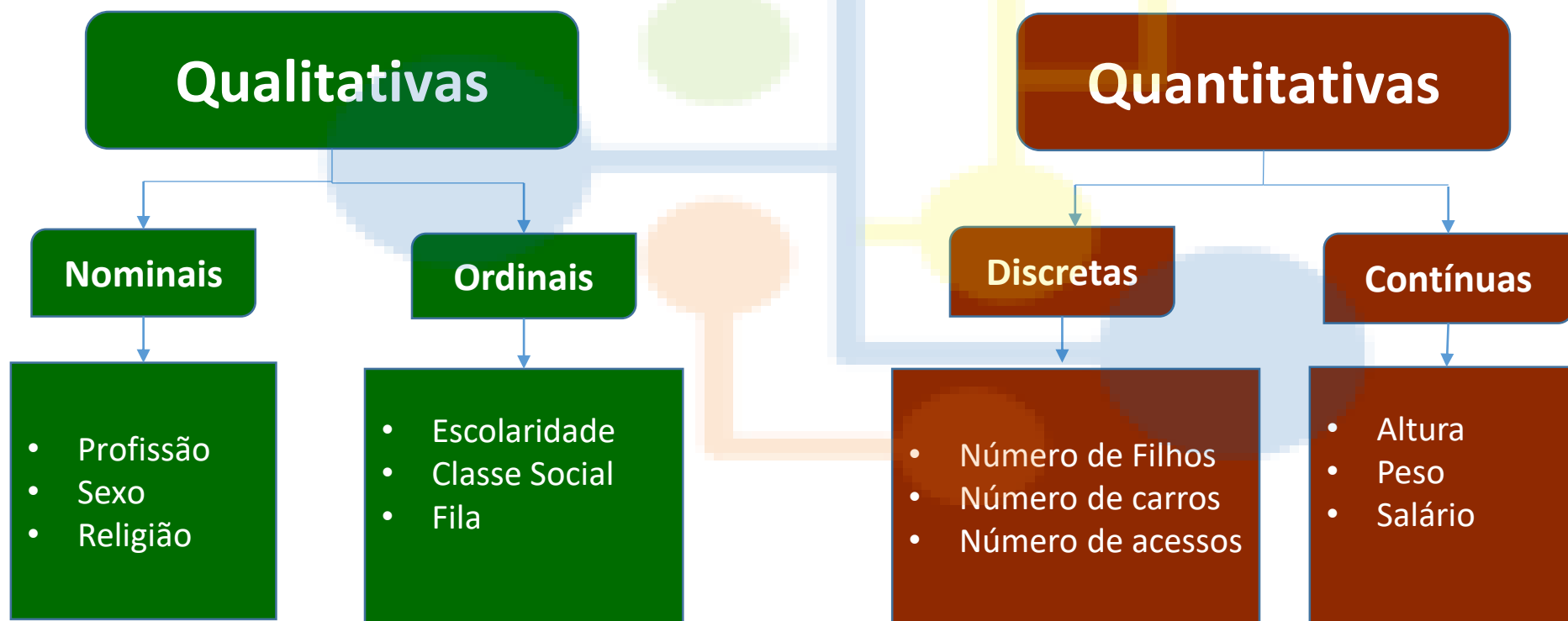
	Idade	Sexo	Peso	Cor dos olhos
Indivíduo 1	42	M	59	Verde
Indivíduo 2	34	M	54	Castanho
Indivíduo 3	56	F	89	Azul
Indivíduo 4	41	M	76	Castanho
Indivíduo 5	23	F	65	Castanho





# Tipos de Variáveis

Dentro desta classificação, podemos ter variáveis:





# Tipos de Variáveis

Um dado classificado como "**idade**" é **quantitativo**

Ex.: 11, 15, 18, 25, 42 anos.

Entretanto, se esse dado for informado por "**faixa etária**" ele é **qualitativo** (ordinal).

Ex: 0 – 5 anos.

6 – 12 anos.

13 – 18 anos.

19 – 28 anos.



# Tipos de Variáveis

É muito importante classificar os tipos de dados das variáveis, pois eles permitirão a você escolher o melhor teste estatístico a ser utilizado na análise dos dados.



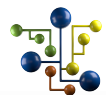
Data Science  
Academy

Data Science Academy adeilsonjunior@gmail.com 5aabf79e5e4cdefd6e8b457e

# Big Data Analytics com R e Microsoft Azure Machine Learning

## Exercício - Colocando os Dados em Contexto

### Seja Bem-Vindo(a)!



# Exercício - Colocando os Dados em Contexto

Uma rede de academias para mulheres decidiu fazer um estudo para verificar a influência do exercício físico praticado por gestantes no peso de seus bebês.



# Exercício - Colocando os Dados em Contexto

Uma rede de academias para mulheres decidiu fazer um estudo para verificar a influência do exercício físico praticado por gestantes no peso de seus bebês.

O objetivo é avaliar a necessidade de mudança ou não no formato dos exercícios oferecidos nas sessões com as futuras mães.



# Exercício - Colocando os Dados em Contexto

## Compreensão do Problema

- Quanto uma grávida costuma se exercitar?
- O grau do exercício influencia o peso do bebê?
- Que pesos são mais comuns para os bebês?



# Exercício - Colocando os Dados em Contexto

Nível de exercício	Peso (em gramas)
nenhum	3242,82
mudando	3547,59
mudando	3929,22
nenhum	2765,92
baixo/moderado	3134,82
mudando	2693,38
mudando	3144,96
nenhum	3508,47
alto	3728,29
nenhum	4012,09
nenhum	3973,98
mudando	3342,50
mudando	3278,79
mudando	3369,27
baixo/moderado	3583,00
nenhum	2323,93







# Exercício - Colocando os Dados em Contexto

Responda:

Pergunta	Resposta
1- Qual a população?	Todas as gestantes que frequentam a rede de academias (500).
2- Qual a amostra?	50 gestantes.
3- Qual a fonte de dados (Primário ou Secundário)?	Primário (você fez a coleta dos dados).
4- Quantas observações e variáveis?	50 observações e 2 variáveis.
5- Quais os tipos de variáveis?	Nível de Exercícios – qualitativa ordinal Peso – quantitative contínua
6- Estes dados ajudam a responder as perguntas ou precisamos de mais dados?	Não. Precisaríamos de dados sobre alimentação, condições de saúde da gestante, tempo total de gravidez, etc...



# Exercício - Colocando os Dados em Contexto

A simples visualização dos dados, ainda que contenha toda a informação, muitas vezes não diz nada.



# Exercício - Colocando os Dados em Contexto

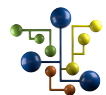
Simplesmente olhar para os dados não fornece um quadro claro do que pode estar acontecendo, especialmente quando a quantidade de dados for muito grande.



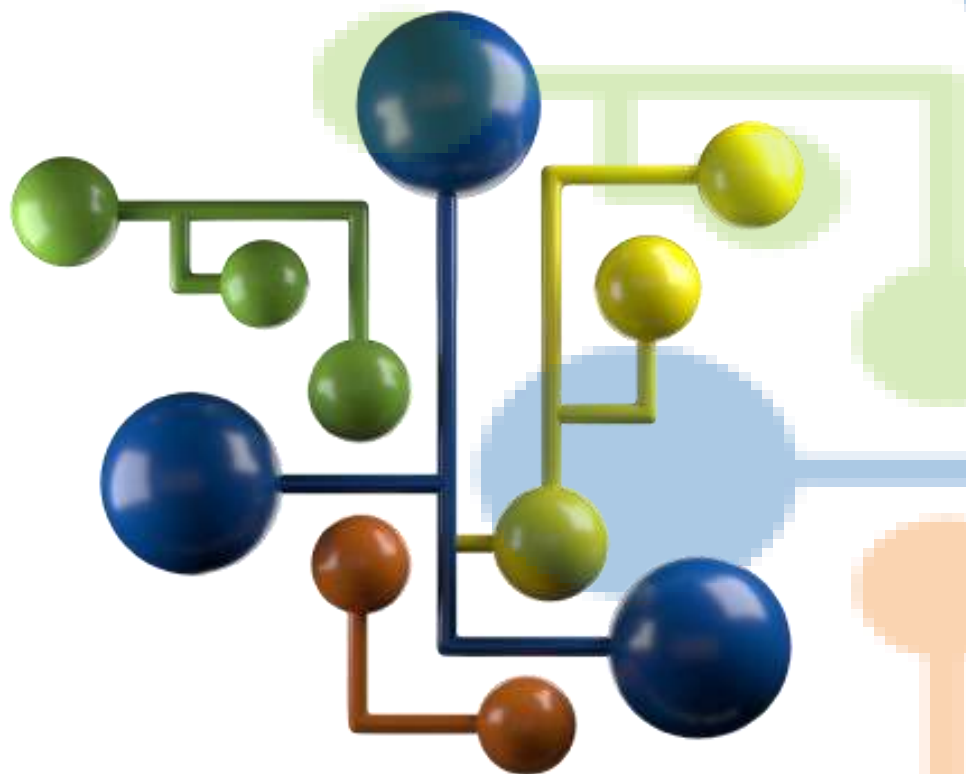
## Exercício - Colocando os Dados em Contexto

Por isso podemos ensinar algoritmos a fazer isso por nós.  
Exatamente onde começa o trabalho em Machine Learning.





# Muito Obrigado por Participar!



Tenha uma Excelente Jornada de Aprendizagem.

Equipe Data Science Academy