

Formação Livre em Ciência de Dados

Aula 2 | Módulo Básico II

João Pedro Passos Pereira

Roteiro Aula 2

1. Parte 1:
 - a. Conceitos básicos em Ciência de Dados I;
2. Parte 2:
 - a. Introdução à Estatística;
 - b. Lógica I.

Parte 1

(1) Conceitos básicos em Ciência de Dados I

Antes de começar...

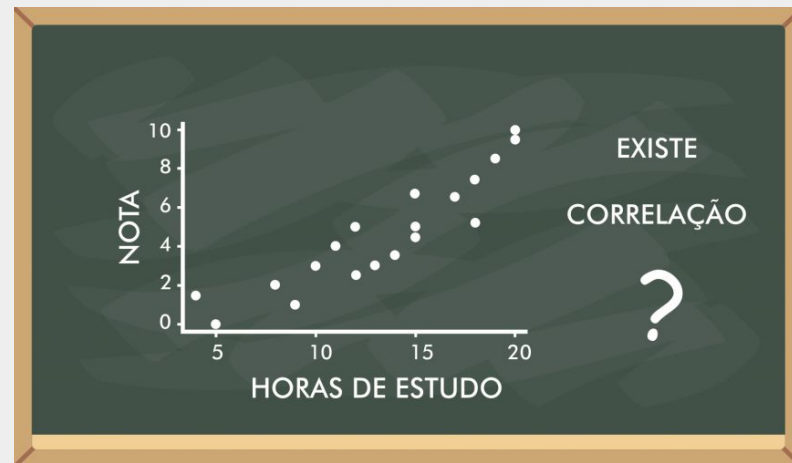
Ideia fundamental

Vimos que a Ciência de Dados é um processo, que transforma dados brutos em dados líquidos (informações), que por sua vez viram conhecimentos e somente por fim inteligência para pessoas físicas, pessoas jurídicas de direito público ou de direito privado. Porém, é necessário entender que:

A Ciência de Dados visa não somente descobrir a existência verdadeira de relações entre dados mas quantificar tais relações (o grau de força envolvido) e explicá-las.

Exemplo

O que poderia quantificar?



Podemos obter um valor “ ρ ” para quantificar esta relação entre variáveis...

Valor de ρ (+ ou -)	Interpretação
0.00 a 0.19	Uma correlação bem fraca
0.20 a 0.39	Uma correlação fraca
0.40 a 0.69	Uma correlação moderada
0.70 a 0.89	Uma correlação forte
0.90 a 1.00	Uma correlação muito forte

E depois classificar este valor de acordo com sua força como na tabela acima

Conceitos básicos (1)



Linguagem R

“R é uma linguagem de programação multi-paradigma orientada a objetos, programação funcional, dinâmica, fracamente tipada(*), voltada à manipulação, análise e visualização de dados.”

“A linguagem R é largamente usada entre estatísticos e analistas de dados para desenvolver software de estatística e análise de dados. Pesquisas e levantamentos com profissionais da área mostram que a popularidade do R aumentou substancialmente nos últimos anos.”

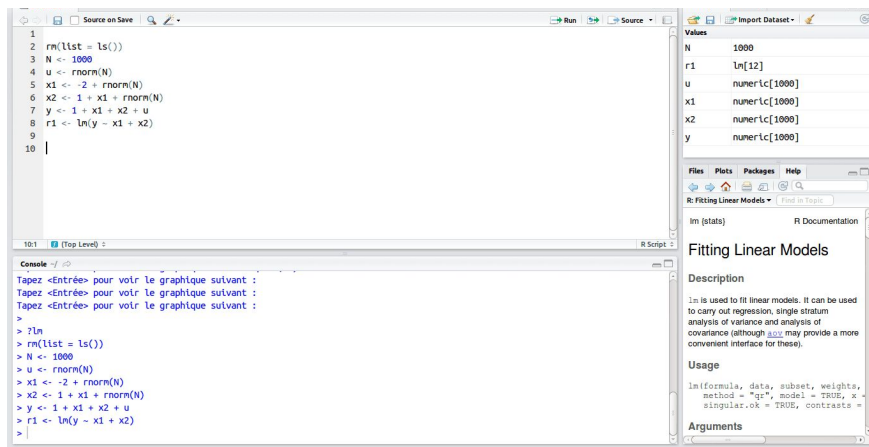
*Usa variáveis com tipos específicos

Conceitos básicos (2)



RStudio

“RStudio é um software livre de ambiente de desenvolvimento integrado para R, uma linguagem de programação para gráficos e cálculos estatísticos.”



Conceitos básicos (3)



Python

“Python é uma linguagem de programação de alto nível, interpretada de script, imperativa, orientada a objetos, funcional (...)”

“Alguns dos maiores projetos que utilizam Python são (...) o site do YouTube e o cliente original do BitTorrent. Grandes organizações que usam a linguagem incluem Google (...), Yahoo! (para o site de grupos de usuários) e NASA. O sistema de gerenciamento de reservas da Air Canada também usa Python em alguns de seus componentes. A linguagem também tem bastante uso na indústria da segurança da informação.”

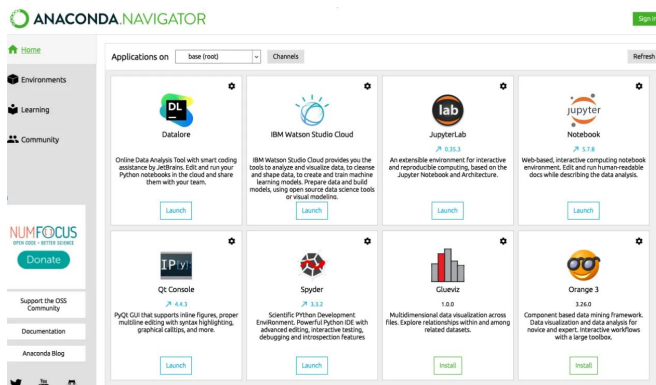
Conceitos básicos (4)



ANACONDA®

Anaconda

“Anaconda é uma distribuição das linguagens de programação Python e R para computação científica (ciência de dados, aplicativos de aprendizado de máquina, processamento de dados em grande escala, análise preditiva, etc.), que visa simplificar o gerenciamento e implantação de pacotes.”

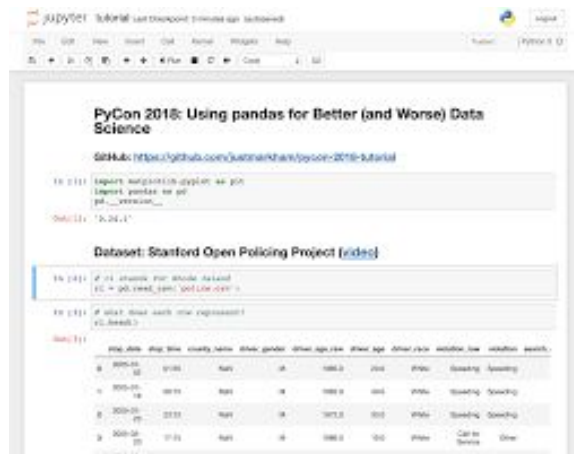


Conceitos básicos (5)



Jupyter Notebook

É um ambiente de desenvolvimento contido na distribuição Anaconda.

A screenshot of a Jupyter Notebook interface in a web browser. The browser title is "jupyter: Tutorial via browser (private ip: localhost)". The notebook content includes a title "PyCon 2018: Using pandas for Better (and Worse) Data Science", a GitHub link, and code cells. The first code cell imports numpy and pandas. The second code cell loads a dataset from a URL. The third code cell displays the first few rows of the dataset as a table.

```
In [1]: import numpy as np
import pandas as pd
pd.__version__

Out[1]: '1.0.1'
```

Dataset: Stanford Open Policing Project [\[video\]](#)

```
In [2]: # all rows for state of indiana
IC = pd.read_csv('police.csv')

In [3]: # select those with row representing
IC.head()
```

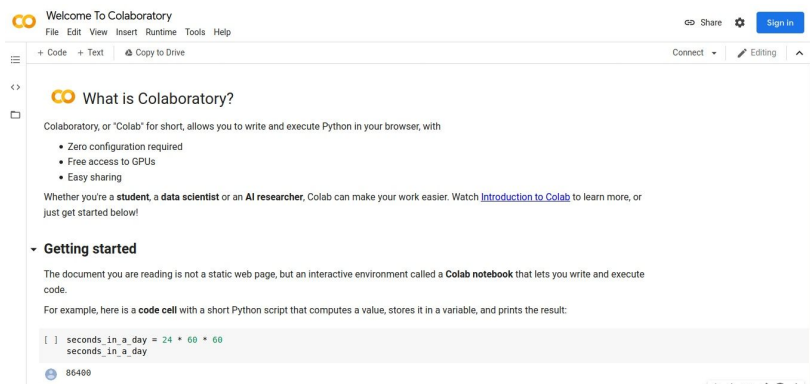
stop_date	stop_time	county_name	driver_gender	driver_age_min	driver_age_max	driver_age	driver_race	violation_low	violation_high	violation
2015-01-01	10:00	harris	M	188.0	200	194.0	Driving	Speeding	Speeding	
2015-01-01	10:00	harris	M	188.0	200	194.0	Driving	Speeding	Speeding	
2015-01-01	10:00	harris	M	188.0	200	194.0	Driving	Speeding	Speeding	
2015-01-01	10:00	harris	M	188.0	200	194.0	Driving	Speeding	Speeding	

Conceitos básicos (5)

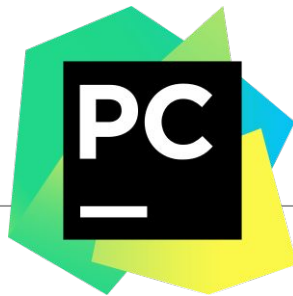


Google Colab

“O Colaboratory ou “Colab” é um produto do Google Research, área de pesquisas científicas do Google. O Colab permite que qualquer pessoa escreva e execute código Python arbitrário pelo navegador e é especialmente adequado para aprendizado de máquina, análise de dados e educação.”



Conceitos básicos (6)



IDEs

“IDE, do inglês Integrated Development Environment ou Ambiente de Desenvolvimento Integrado, é um programa de computador que reúne características e ferramentas de apoio ao desenvolvimento de software com o objetivo de agilizar este processo.”

Ou seja, são outros ambientes de desenvolvimento, alternativos aos mencionados. Há inúmeros. Obs: Sublime Text não é um ambiente completo (IDE), mas substitui bem e por isso é considerado.



Conceitos básicos (7) - Parte 1

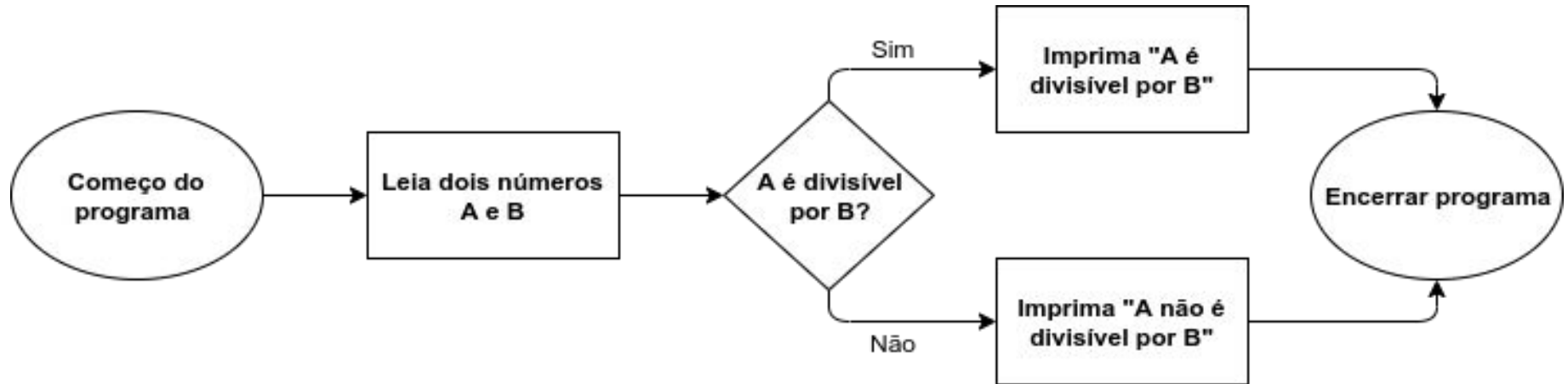
Programação orientada a objetos (POO)

“Como a maioria das atividades que fazemos no dia a dia, programar também possui modos diferentes de se fazer. Esses modos são chamados de paradigmas de programação e, entre eles, estão a **programação orientada a objetos (POO)** e a **programação estruturada**.”

“Na programação estruturada, um programa é composto por três tipos básicos de estruturas:

- sequências: são os comandos a serem executados
- condições: sequências que só devem ser executadas se uma condição for satisfeita (exemplos: if-else e comandos parecidos)
- repetições: sequências que devem ser executadas repetidamente até uma condição for satisfeita (for, while, do-while etc)”

Linguagem orientada a objetos (POO) - Parte 1.1



Conceitos básicos (7) - Parte 2

Linguagem orientada a objetos (POO)

“(...) o acesso às variáveis não possui muitas restrições na programação estruturada. Em linguagens fortemente baseadas nesse paradigma, restringir o acesso à uma variável se limita a dizer se ela é visível ou não dentro de uma função (ou módulo, como no uso da palavra-chave static, na linguagem C), **mas não se consegue dizer de forma nativa que uma variável pode ser acessada por apenas algumas rotinas do programa.** O contorno para situações como essas envolve práticas de programação danosas ao desenvolvimento do sistema, como o uso excessivo de variáveis globais. Vale lembrar que variáveis globais são usadas tipicamente para manter estados no programa, marcando em qual parte dele a execução se encontra.”

Conceitos básicos (7) - Parte 2.1

Programação orientada a objetos (POO) - O que é uma variável na programação?

“Na programação, uma variável é um **objeto capaz de reter e representar um valor ou expressão**. Enquanto as variáveis só "existem" em tempo de execução, elas são associadas a "nomes", chamados identificadores, durante o tempo de desenvolvimento.”

“Quando nos referimos à variável, do ponto de vista da programação de computadores, estamos tratando de uma “região de memória (do computador) previamente identificada cuja finalidade é armazenar os dados ou informações de um programa por um determinado espaço de tempo”. **A memória do computador se organiza tal qual um armário com várias divisões. Sendo cada divisão identificada por um endereço diferente em uma linguagem que o computador entende.**”

Conceitos básicos (7) - Parte 3

Programação orientada a objetos (POO)

Então, o que de fato é a POO?

“A programação orientada a objetos surgiu como uma alternativa a essas características da programação estruturada. O intuito da sua criação também foi o de **aproximar o manuseio das estruturas de um programa ao manuseio das coisas do mundo real**, daí o nome "objeto" como uma algo genérico, que pode representar qualquer coisa tangível.”

“As duas bases da POO são os conceitos de **classe e objeto**. Desses conceitos, derivam alguns outros conceitos extremamente importantes ao paradigma, que não só o definem como são as soluções de alguns problemas da programação estruturada. **Os conceitos em questão são o encapsulamento, a herança, as interfaces e o polimorfismo.**”

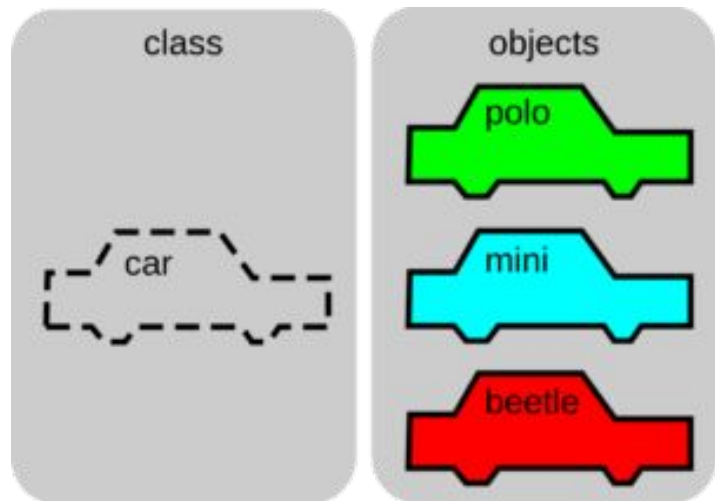
Conceitos básicos (7) - Parte 3.1

Classe e Objeto

“Uma classe é um conjunto de **características e comportamentos** que definem o conjunto de objetos pertencentes à essa classe”

“Repare que a classe em si é um **conceito abstrato**, como um molde, que se torna concreto e palpável através da criação de um objeto.”

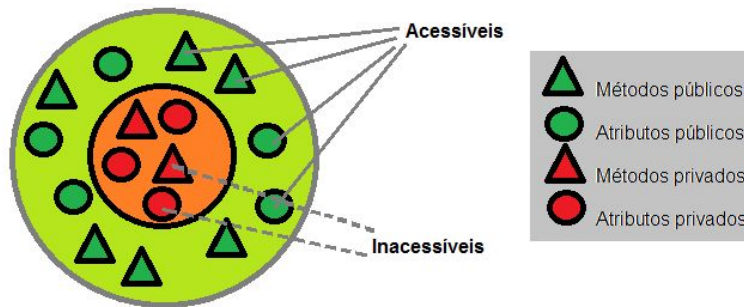
“Chamamos essa criação de **instanciação da classe**, como se estivéssemos usando esse **molde** (classe) para criar um objeto.”



Conceitos básicos (7) - Parte 3.2

Encapsulamento

“Encapsulamento é a técnica que faz com que detalhes internos do funcionamento dos métodos de uma classe permaneçam ocultos para os objetos. Por conta dessa técnica, o conhecimento a respeito da implementação interna da classe é desnecessário do ponto de vista do objeto, uma vez que isso passa a ser responsabilidade dos métodos internos da classe.”



Conceitos básicos (7) - Parte 3.2

```
class Funcionario:
    def __init__(self, nome, cargo, valor_hora_trabalhada):
        self.nome = nome
        self.cargo = cargo
        self.valor_hora_trabalhada = valor_hora_trabalhada
        self.__salario = 0
        self.__horas_trabalhadas = 0

    @property
    def salario(self):
        return self.__salario

    @salario.setter
    def salario(self, novo_salario):
        raise ValueError("Impossível alterar salario diretamente. Use a funcao calcula_salario().")

    def registra_hora_trabalhada(self):
        self.__horas_trabalhadas += 1

    def calcula_salario(self):
        self.__salario = self.__horas_trabalhadas * self.valor_hora_trabalhada

pedro = Funcionario('Pedro', 'Gerente de Vendas', 50)
pedro.salario = 100000
```

Conceitos básicos (7) - Parte 3.2

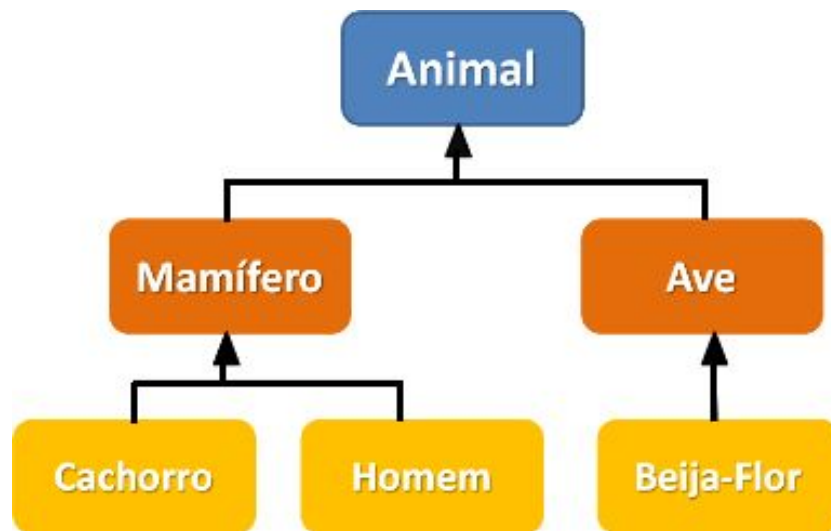
Em Python: “A função `init`, também chamada de “método construtor”, possui a responsabilidade de criar o objeto daquela classe. Nem sempre você precisará criá-la, porém, se o seu projeto exige que você utilize essa função, nela será contida todas as informações principais do objeto. Um exemplo disso pode ser o objeto “Pessoa”.”

```
class Pessoa:
    def __init__(self, nome, idade):
        self.nome = nome
        self.idade = idade
```

Conceitos básicos (8) - Parte 3.2

Herança

“Quando dizemos que uma classe A é um tipo de classe B, dizemos que a classe A herda as características da classe B e que a classe B é mãe da classe A, estabelecendo então uma relação de herança entre elas.”



Conceitos básicos (7) - Parte 3.2

Em Python: “O `super()` é utilizado entre heranças de classes, ele nos proporciona estender/subscrever métodos de uma super classe (classe pai) para uma sub classe (classe filha), através dele definimos um novo comportamento para um determinado método construído na classe pai e herdado pela classe filha.”

```
class Veiculo:
    def __init__(self, tipo, chassi, marca, modelo, ano):
        self.tipo = tipo
        self.chassi = chassi
        self.marca = marca
        self.modelo = modelo
        self.ano = ano

class Motocicleta(Veiculo):
    def __init__(self, tipo, chassi, marca, modelo, ano, cilindrada):
        super().__init__(tipo, chassi, marca, modelo, ano)
        self.cilindrada = cilindrada

x = Motocicleta(cilindrada=3, tipo=0, chassi=0, marca=0, modelo=0, ano=0)
print(x.cilindrada)
```


Parte 2

(2) Introdução à Estatística

“Estatística é um conjunto de métodos e processos quantitativos que serve para estudar e medir os fenômenos coletivos.”

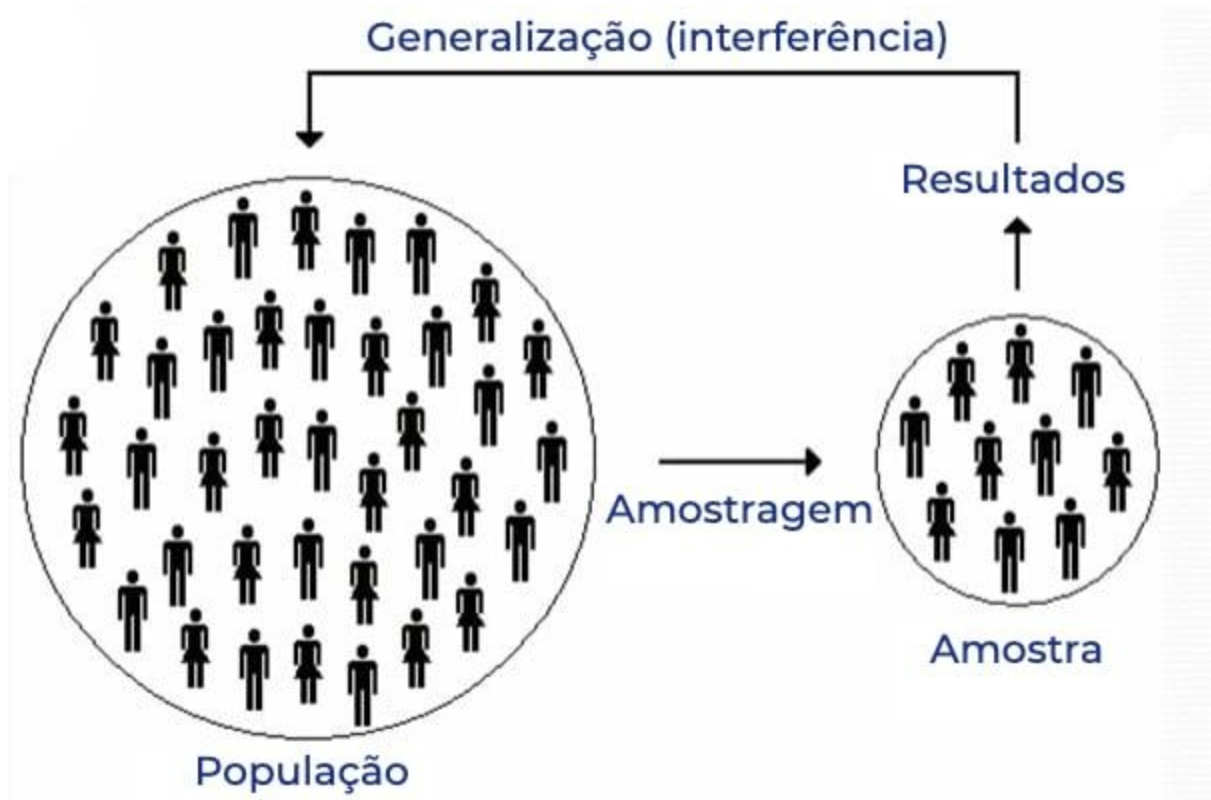
Estatística para os cursos de Economia, Administração e Ciências Contábeis
3ª edição - 2009

Conceitos iniciais (1)

População	Amostra	Parâmetro	Estimador
“Conceituaremos População como sendo o conjunto de todos os itens (pessoas, coisas, objetos) que interessam ao estudo de um fenômeno coletivo segundo alguma característica. ”	“Entenderemos por amostra, qualquer subconjunto não vazio de uma população” E podemos acrescentar que é um subconjunto menor que a População.	“Uma característica numérica estabelecida para toda uma população é denominada <u>parâmetro</u> .“	“Uma característica numérica estabelecida para uma amostra é denominada <u>estimador</u> .”

Conceitos iniciais (2)

Estatística Descritiva	Estatística Inferencial ou Indutiva
<p data-bbox="117 467 977 549">“É a parte da Estatística que tem por objeto descrever os dados observados.”</p> <p data-bbox="117 603 956 691">A Estatística Descritiva, na sua função de descrição dos dados, tem as seguintes atribuições:</p> <ul data-bbox="117 696 933 964" style="list-style-type: none">a) A obtenção dos dados estatísticos;b) A organização dos dados;c) A redução dos dados;d) A representação dos dados;e) A obtenção de algumas informações que auxiliam a descrição do fenômeno observado.	<p data-bbox="1047 467 1850 647">“É a parte da Estatística que tem por objetivo obter e generalizar conclusões para a população a partir de uma amostra, através do cálculo de probabilidade.</p> <p data-bbox="1047 696 1850 778">O cálculo de probabilidade é que viabiliza a inferência estatística”</p>



Conceitos iniciais (3)

Censo	Estimação
<p>“É uma avaliação direta de um parâmetro, utilizando-se todos os componentes da população.”</p> <p>Propriedades Principais do Censo:</p> <ol style="list-style-type: none">1. Admite erro processual zero e tem confiabilidade 100%;2. É caro;3. É lento;4. É quase sempre desatualizado;5. Nem sempre é viável.	<p>“É uma avaliação indireta de um parâmetro, com base em um estimador através do cálculo de probabilidades.”</p> <p>Propriedades Principais da Estimação:</p> <ol style="list-style-type: none">1. Admite erro processual positivo e tem confiabilidade menor que 100%;2. É barata;3. É rápida;4. É atualizada;5. É sempre viável.

Quem visitará o domicílio?



CENSO 2022

Recenseador
e/ou supervisor

PNAD CONTÍNUA

Agente de pesquisa
e/ou supervisor



Quando será a pesquisa?

CENSO 2022

1º de agosto a
dezembro de 2022



PNAD CONTÍNUA

A PNAD Contínua é uma pesquisa que visita seu domicílio cinco vezes – uma vez a cada três meses – durante cinco trimestres consecutivos.



Conceitos iniciais (4)

Dados Brutos (Ciência de Dados)	Dados Brutos (Estatística)
Dado antes de ser separado por relevância para o processo que será iniciado. Ou seja, dado derivado de coleta, sem classificação de relevância .	Sequência de valores obtidos pela observação de um fenômeno, não ordenados, manipulados ou analisados. Obs: não há preocupação efetiva com a relevância do dado no contexto uma vez a que a estatística é vista como uma ferramenta .

Conceitos iniciais (5)

Rol

Quando se ordena os Dados Brutos estatísticos em uma sequência crescente ou decrescente, temos a existência de um ROL.

X: 4; 8; 7; 5; 6; 5 (Dados Brutos)

OU

X: 4; 5; 5; 6; 7; 8 (ROL)

Conceitos iniciais (6)

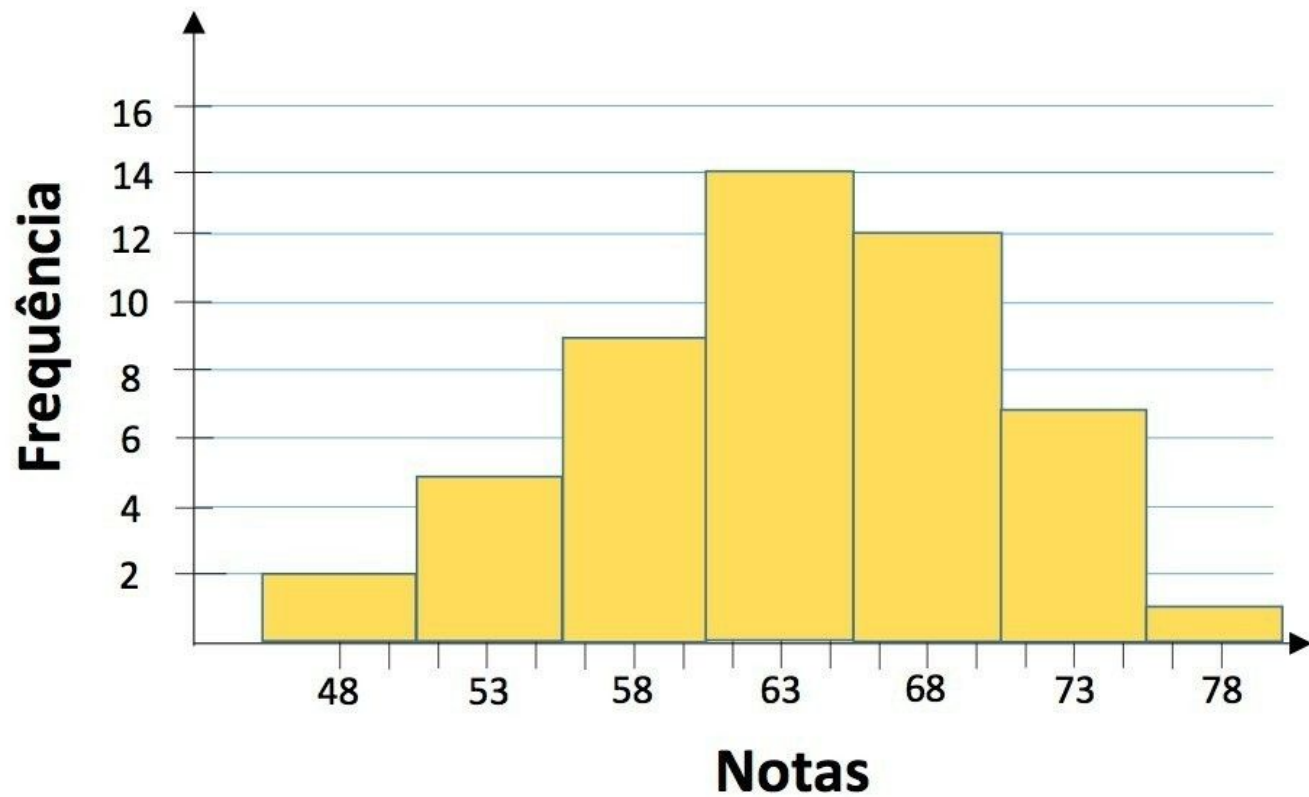
Variável

“Em estatística, uma variável é a **característica** dos elementos da amostra ou população que nos interessa averiguar estatisticamente. Uma variável estatística é uma **característica** que admite **diferentes valores**, um por cada unidade estatística.”

Conceitos iniciais (7)



Variável quantitativa	Variável qualitativa
<p>“São aquelas cujos valores são expressos em números.”</p> <p>Tipos:</p> <ul style="list-style-type: none">- <u>Discreta</u>: números inteiros (resultado finito, enumerável)- <u>Contínua</u>: Números do conjunto dos números reais	<p>"São aquelas cujos valores podem ser separados em diferentes categorias que se distinguem por alguma característica não numérica. Por exemplo: sexo (masculino e feminino), cor dos cabelos (preto, loiro, ruivo, castanho, etc)."</p> <p>Tipos:</p> <ul style="list-style-type: none">- <u>Nominal</u> (em que a ordem não tem importância)- <u>Ordinal</u> (em que a ordem assume papel relevante)



(3) Lógica I

Conceitos iniciais (1)

Argumento	Sentença declarativa
“Sucintamente, um argumento é um conjunto de sentenças declarativas ou enunciados , que, por sua vez, são aquelas que admitem um valor-verdade e podem ser verdadeiras ou falsas ”	<p>Segundo Aristóteles, existem dois tipos de sentenças declarativas:</p> <ol style="list-style-type: none">1. As que relacionam objetos a categorias:<ol style="list-style-type: none">1. x é P — x é um objeto e P uma categoria.2. x não é P2. As que relacionam categorias a categorias:<ol style="list-style-type: none">1. Todo P é Q — universal positiva2. Nenhum P é Q — universal negativa3. Algum P é Q — existencial positiva4. Algum P não é Q — existencial negativa

Conceitos iniciais (2)

Validez	Consistência
<p>Além dos elementos básicos de um argumento definidos por Aristóteles, surge a necessidade de caracterizá-lo como algo não apenas descritivo. Isso é possível ao se verificar que um argumento envolve um ou mais atos de inferência a partir de sentenças tomadas como premissas, que permitem tirar conclusões.</p> <p>Premissa 1 Premissa 2 ... Premissa n logo, Conclusão</p> <p>Dizemos que um ato de inferência é válido se se todas as suas premissas forem verdadeiras, a conclusão também for. Quando isso acontece, dá-se a ele o nome de <u>silogismo</u>.</p>	<p>Um conjunto de sentenças é dito inconsistente se existe um ato de inferência baseado em algum subconjunto desse conjunto que deduz validamente uma sentença e, com base em outro subconjunto desse conjunto, deduz validamente uma sentença contraditória a anterior. O conjunto é consistente caso contrário.</p> <p>Ex: Todo A é B. Todo B é D. Algum A não é D.</p> <p>Resultado: Falso</p>

Conceitos iniciais (2)

Ex: Todo A é B. Todo B é D. Algum A não é D.

Se todo A é B, temos que $A \subseteq B$. Se todo B é D, temos que $B \subseteq D$. Por propriedade de conjuntos, concluímos seguramente que $A \subseteq D$. Porém, temos uma terceira sentença, "Algum A não é D", que nos diz que $A \not\subseteq D$. Como $A \subseteq D$ e $A \not\subseteq D$ é impossível, temos uma contradição. Quando isso ocorre, dizemos que o conjunto de sentenças é inconsistente.

Contatos

Meus links úteis

Instagram: @j0pewd2

Site: <https://joaopedropereira.com.br>

E-mail: contato@joaopedropereira.com.br

Linkedin: <https://linkedin.com/in/joaopedrowd/>