Ciência de Dados - Princípios e Aplicações

Prof. Eng. Ranyeri do Lago Rocha
e-mail ranyeri.rocha@inatel.br

Agenda

- Relações entre Ciência de Dados, Aprendizado de Máquina e Probabilidade e Estatística
- Análise de Dados

Relações entre Ciência de Dados, Aprendizado de Máquina e Probabilidade e Estatística

Relações entre Ciência de Dados, Aprendizado de Máquina e Probabilidade e Estatística

- Probabilidade e Estatística
- Aprendizado de Máquina

Probabilidade e Estatística

Estatística

O termo pode se referir a:

- Estatística Consolidada: cálculo de valores numéricos particulares de interesse a partir dos dados. São exemplo destes valores: somas, médias, taxas, etc. Ainda, a estatística consolidada pode ser condicionalmente calculada, para uma parte da população. *Atenção adicional para a distribuição dos dados!*
- Estatística: compreensão da distribuição dos dados, quais estatísticas são apropriadas, compreender como usar os dados para testar hipóteses e para estimar a incerteza de conclusões. Além, bastante relacionada, está a quantificação da incerteza em intervalos de confiança.

Probabilidade e Estatística

Estatística

Refere-se ao uso de matemática e técnicas que podemos entender os dados.
Como descrever os dados? A estatística apresenta elementos essenciais!

Para conjuntos de dados pequenos o suficiente, o próprio dataset é a melhor descrição.
Para conjuntos de dados grandes, olhar para o conjuntos não é muito informativo.

- → Histograma
- → Quantidade de dados
- → Menor e maior valor
- → Tendências
  - → Média e Mediana
  - → Quantís (Quartís ou outros valores quaisquer)
- → Dispersão
  - → Variância (da média) e desvio padrão
- → Correlação
  - → Covariância (de outra variável)
    - versão normalizada da covariância, chamada Coeficiente de Correlação.

Probabilidade e Estatística

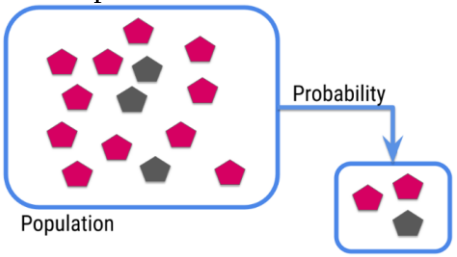
Probabilidade

Pensamos aqui sobre a probabilidade de um evento acontecer baseado em um universo de possibilidades. Na ciência de dados, a probabilidade é usada para construir e avaliar modelos.

→ Teorema de Bayes (estimativa de probabilidade de eventos)
→ Probabilidade associada a modelos de aprendizado de máquina, principalmente

Probabilidade e Estatística

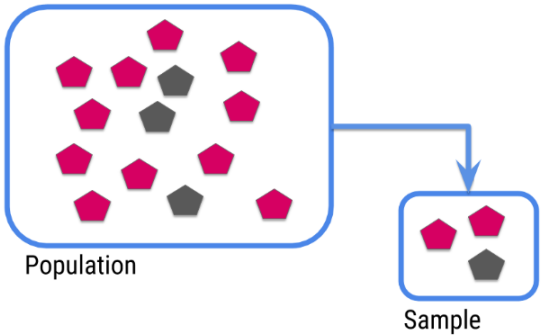
É provável?



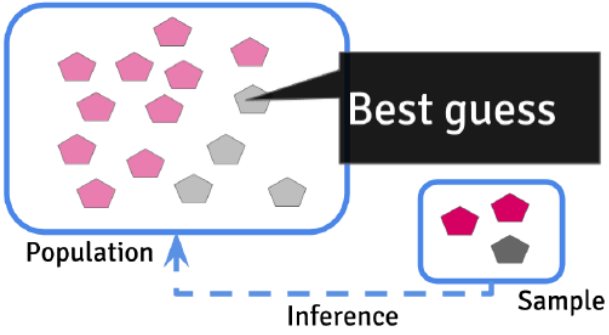
Análise Inferencial

- Gerando uma estimativa e medindo a incerteza sobre esta estimativa

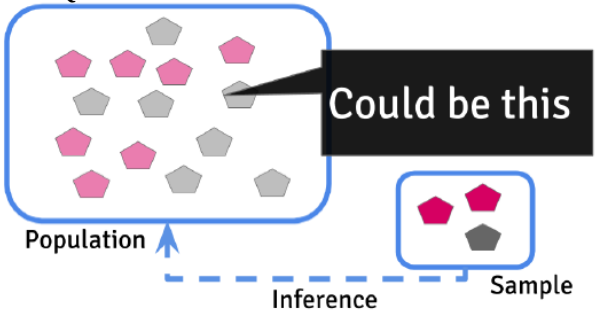
Obter amostragem



Melhor **estimativa**



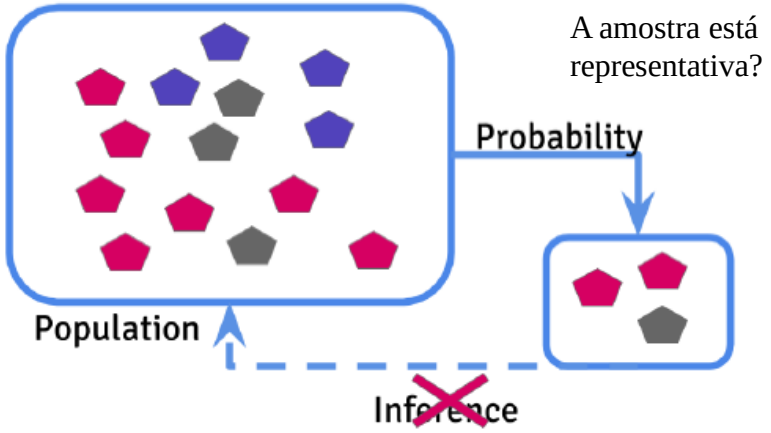
Qual a incerteza?



Probabilidade e Estatística

Probabilidade

- Inferência depende muito de como o dado foi amostrado
- Se possível, faça uma análise exploratória e confirme em conjunto de dados separados
- Tenha certeza e cuidado ao escolher a população, a amostra para a análise. Se não for representativa do universo real, os resultados serão tendenciosos (*biased*)



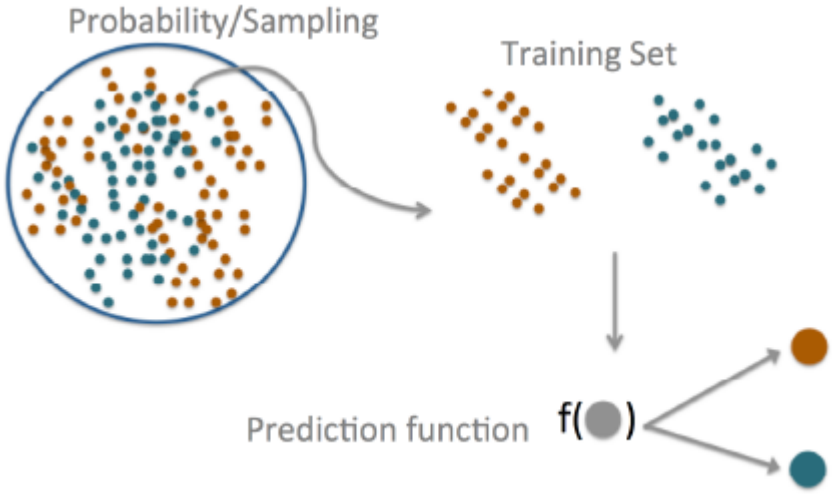
Relações entre Ciência de Dados, Aprendizado de Máquina e Probabilidade e Estatística

- Probabilidade e Estatística
- Aprendizado de Máquina

Aprendizado de Máquina

Apesar de ser um senso comum achar que Cientista de Dados usa praticamente só modelos de aprendizado de máquina no seu trabalho diário, a verdade é que o real trabalho em Ciência de Dados é transformar problemas de negócio em problemas de dados e:

1. coletar dados
2. entender os dados
3. limpar os dados
4. formatar os dados
5. Se preciso, usar aprendizado de máquina.



Aprendizado de Máquina

Modelo x Aprendizado de Máquina

Modelo é uma especificação de uma relação matemática ou probabilística que existe entre diferentes variáveis.

Exemplo:
- Melhorar o dinheiro para seu site de rede social
  - Número de usuários
  - Receita por usuário
  - Número de funcionários

→ Lucro anual para os próximos anos

Este modelo pode ser aprendido através dos dados, usando Aprendizado de Máquina!

Regressão

Classificação

Clusterização

Redes Neurais

Aprendizado Profundo

Aprendizado de Máquina

Quase um *check list*

- Divida os dados em conjunto de treinamento e teste (70/30)
- Identificar razões da sua amostra não ser representativa da população
- Mais dados geralmente são melhores que bons algoritmos
- *Features* são mais importantes que o algoritmo
- Defina o erro do modelo antes de começar
    - Matriz de confusão ou MSE?
- Evite *overfitting* com validação cruzada
- Faça predições médias baseada em vários modelos (*ensemble*)
- Faça escolhas:
    - Interpretabilidade vs acurácia
    - Velocidade vs acurácia
    - Simplicidade vs acurácia
    - Escalabilidade vs acurácia

Análise de Dados

Análise de Dados

- Modelagem e Análise de Dados
- Estatística aplicada a análise de dados

Modelagem e Análise de Dados

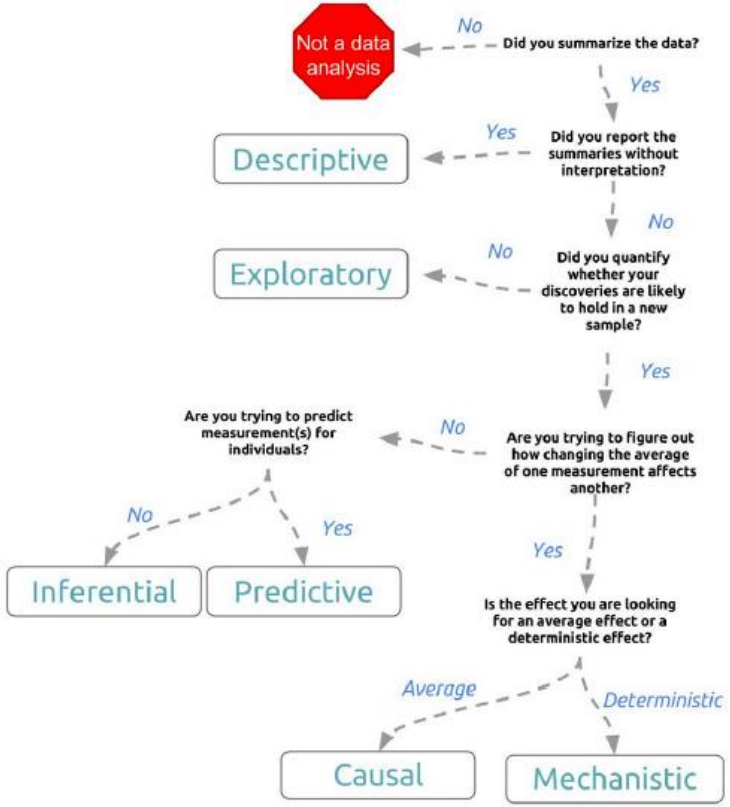
As questões envolvidas no fluxo de análise de dados.

Descritiva: a interpretação não faz parte da análise.

- Entender os componentes de um dataset, descrevê-los, e explicar a descrição para outros que possam querer entender o dataset.
- O Censo, quando faz a coleta dos dados de residência, idade, sexo, etc, tem a intenção de resumir o contexto da população em diferentes partes do país.

Exploratória: construir uma análise procurando descobertas, tendências, correlações, para gerar ideias ou hipóteses.

- Encontrar relacionamentos desconhecidos entre variáveis.



Modelagem e Análise de Dados

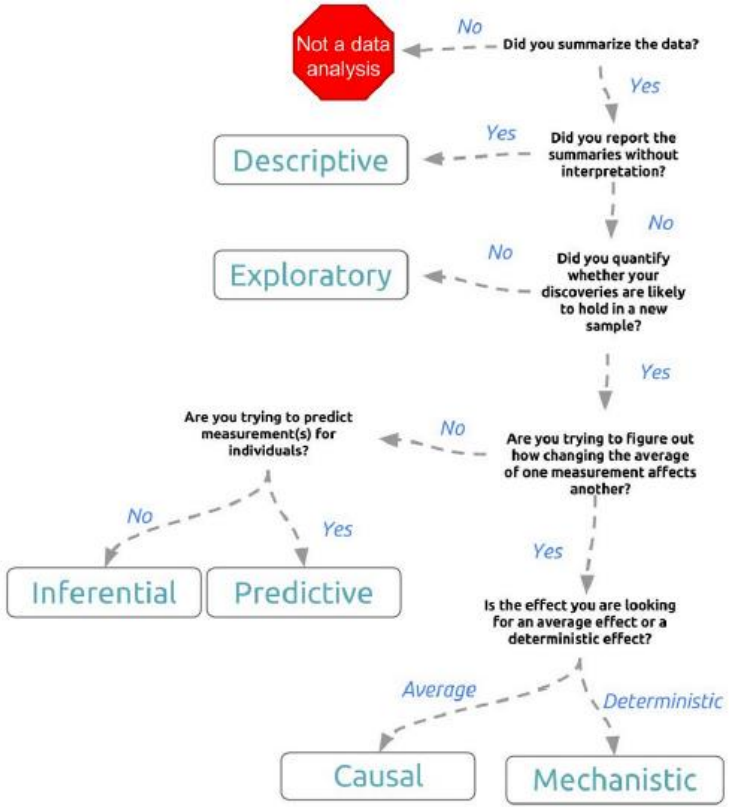
As questões envolvidas no fluxo de análise de dados.

Inferencial: vai além da exploratória. Quantifica se um padrão observado é válido fora do conjunto de dados atual.

- Amostra pequena para dizer algo em novas amostras. Útil quando é custoso obter dados de toda população.

Preditiva: visa extrapolar a análise, usando algumas medidas (features) para predizer outra medida (efeito).
A análise mostra que pode predizer, mas não explica co

- Grande coleção de dados e predição para novos "indivíduos".
Olhar para o futuro ou para características difíceis de medir.



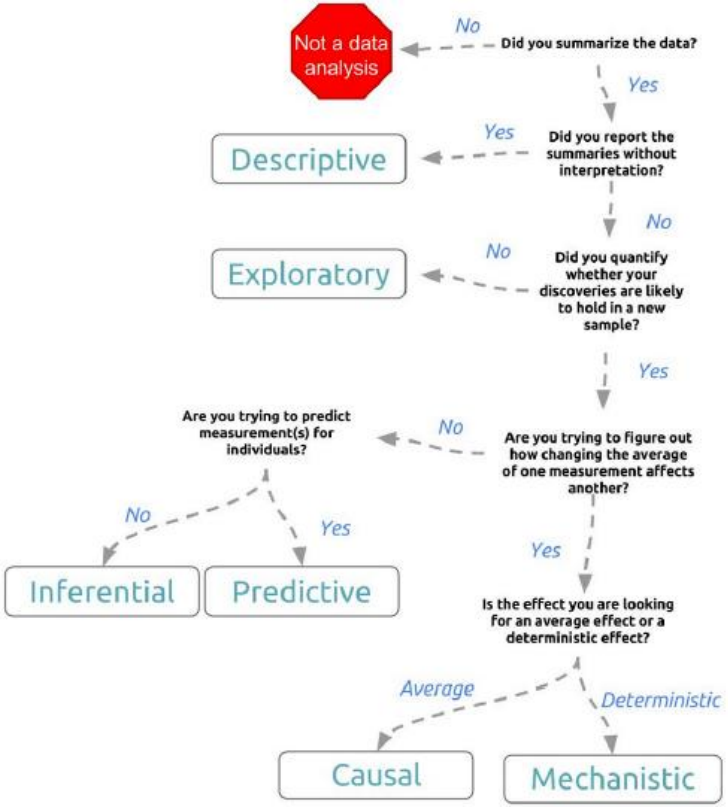
Modelagem e Análise de Dados

As questões envolvidas no fluxo de análise de dados.

Causal: análise que busca o que acontece a uma medida se outra medida de interesse sofrer uma mudança. Causa e efeito.

- A análise causal identifica claramente as direções e magnitudes das relações

Mecanicista: busca efeitos médios em dados ruidosos. A análise busca demostrar que mudanças em uma medida leva sempre e exclusivamente a um comportamento determinístico em outra medida.



Modelagem e Análise de Dados

Uma imagem vale mais que mil palavras

Visualização de Dados fornece uma poderosa forma para comunicar uma descoberta baseada em dados. Em muitos casos, a visualização dos dados é tão convincente que não é preciso análise adicional.

Visualização de Dados é a ferramenta mais forte dentro do que chamamos de **EDA – Exploratory Data Analysis**.

> “O maior valor de uma imagem é quando esta nos força a notar o que nunca esperávamos ver.”

EDA é talvez a parte mais importante da Análise de Dados.

Modelagem e Análise de Dados

EDA – Exploratory Data Analysis

É um importante passo no processo de Ciência de Dados.

- Entender melhor o conjunto de dados
- Checar os atributos e formato
- Validar algumas hipóteses
- Ter uma ideia preliminar sobre o próximo passo a ser tomado

Modelagem e Análise de Dados

Distribuição dos Dados

Tipos de variáveis

- Categóricas
  - Ordinais (ordenados)
  -- (Fraco, Regular, Bom, Ótimo)
  - Não ordinais
  -- (Masculino e Feminino)
- Numéricas
  - Contínuas
  -- Podem assumir qualquer valor dentro de um intervalo
  - Discretas
  -- Assumem apenas valores arredondados (inteiros)

Modelagem e Análise de Dados

Distribuição dos Dados

Imaginem o seguinte:

- Temos informações de altura de homens e mulheres, dados fictícios, para explicar para alguém de outra cultura e país, como somos fisicamente.

Considere os dados estão divididos em: Feminino e Masculino

1. Considerar uma proporção simples entre as categorias
   77% de homens e 23% de mulheres

Modelagem e Análise de Dados

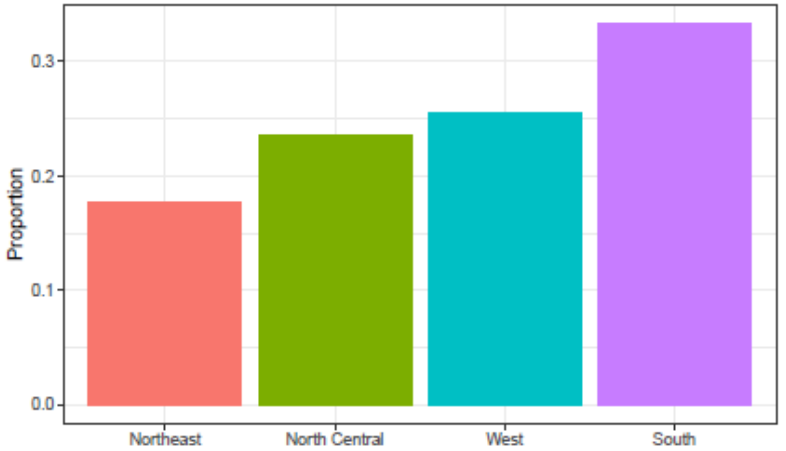
Distribuição dos Dados

Imaginem o seguinte:

- Temos informações de altura de homens e mulheres, dados fictícios, para explicar para alguém de outra cultura e país, como somos fisicamente.

Considere que os dados estão divididos em:
Nordeste, Centro Norte, Oeste e Sul

1. Considerar um gráfico de barras é geralmente suficiente para esta tarefa



Modelagem e Análise de Dados

Distribuição dos Dados

Imaginem o seguinte:

- Temos informações de altura de homens e mulheres, dados fictícios, para explicar para alguém de outra cultura e país, como somos fisicamente.

Considere que os dados estão divididos em:
Altura, em centímetros (valores numéricos)

1. Não é interessante considerar a frequência de cada valor de entrada. Muitos elementos podem ter apenas 1 ocorrência enquanto outros, muitas frequências.
2. Uma forma representativa para estes casos é a chamada **CDF – Cumulative Distribution Function**

Modelagem e Análise de Dados

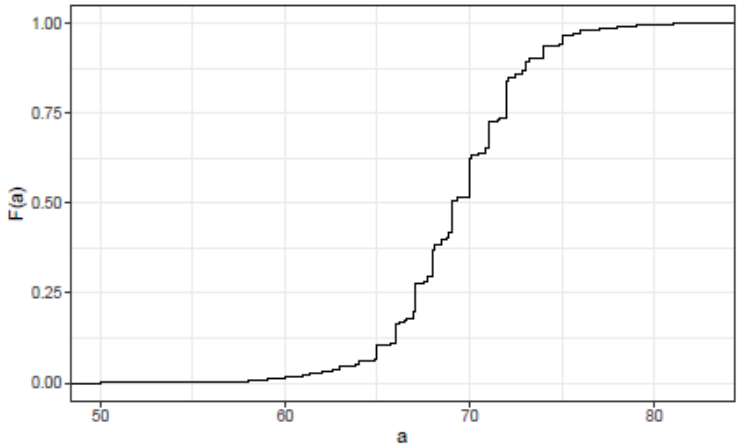
Distribuição dos Dados

Imaginem o seguinte:

- Temos informações de altura de homens e mulheres, dados fictícios, para explicar para alguém de outra cultura e país, como somos fisicamente.

Considere que os dados estão divididos em:
Altura, em centímetros (valores numéricos)

1. A CDF (Cumulative Distribution Function) define a distribuição para dados numéricos.
2. No exemplo, 16% dos valores estão abaixo de 65...
3. No exemplo, 62.5% estão abaixo de 70...



Modelagem e Análise de Dados

Distribuição dos Dados

Imaginem o seguinte:

- Temos informações de altura de homens e mulheres, dados fictícios, para explicar para alguém de outra cultura e país, como somos fisicamente.

Considere que os dados estão divididos em:
Altura, em centímetros (valores numéricos)

1. CDF não é muito utilizado. Qual é o valor de centro dos dados? Os dados possuem uma distribuição simétrica? Quais faixas contém 95% dos valores?
2. Uma boa prática, geralmente preferida, são os histogramas.

Modelagem e Análise de Dados

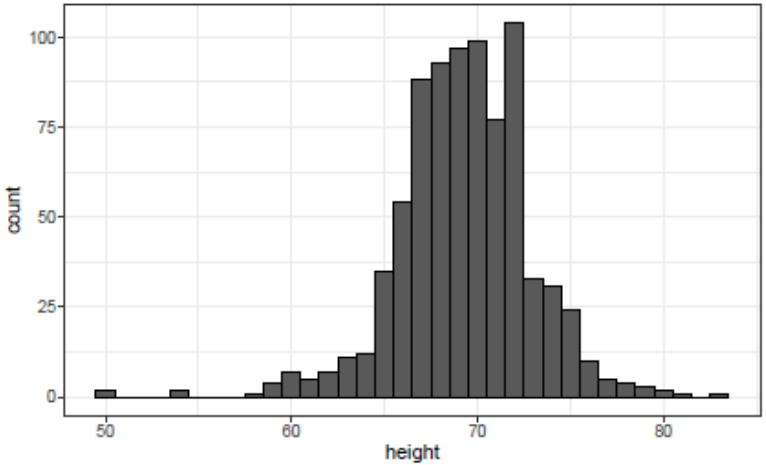
Distribuição dos Dados

Imaginem o seguinte:

- Temos informações de altura de homens e mulheres, dados fictícios, para explicar para alguém de outra cultura e país, como somos fisicamente.

Considere que os dados estão divididos em:
Altura, em centímetros (valores numéricos)

1. Histogramas produzem gráficos que são muito mais fáceis de interpretar.
2. Parece um gráfico de barras, mas, o eixo X é numérico
3. Faixa de valores (50 a 84), a maioria entre 63 e 75, etc...
4. Dados entre as faixas (bins) são negligenciados.



Modelagem e Análise de Dados

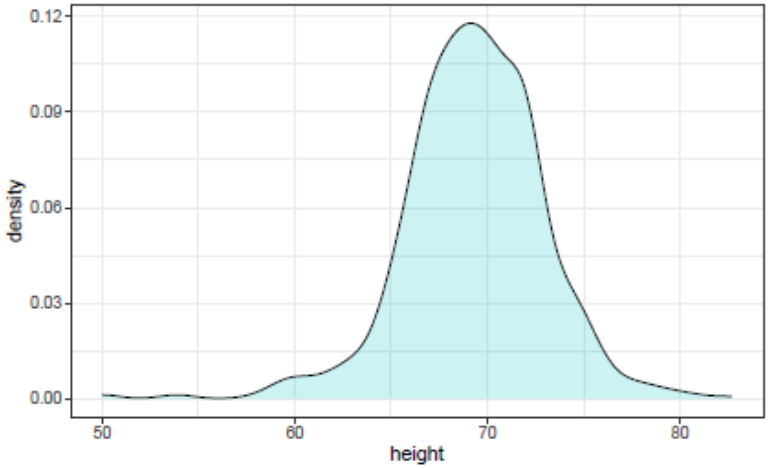
Distribuição dos Dados

Imaginem o seguinte:

- Temos informações de altura de homens e mulheres, dados fictícios, para explicar para alguém de outra cultura e país, como somos fisicamente.

Considere que os dados estão divididos em:
Altura, em centímetros (valores numéricos)

1. Para tentar resolver a negligência do histograma, o gráfico *smoothed density* é apresentado. O nome oficial é KDE (Kernel Density Estimate).



Modelagem e Análise de Dados

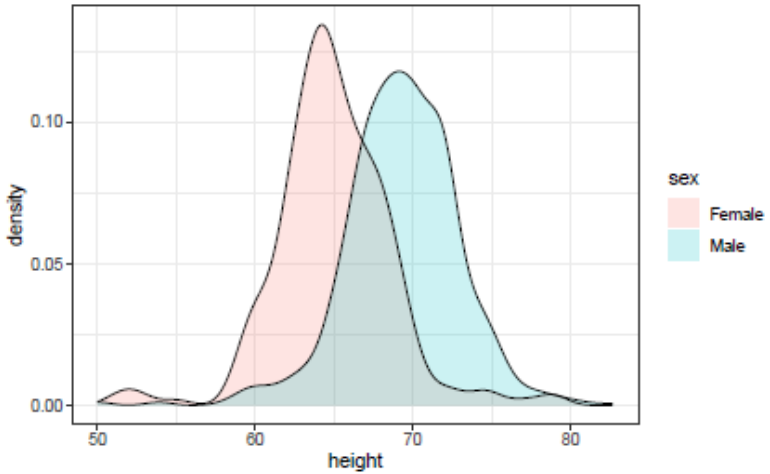
Distribuição dos Dados

Imaginem o seguinte:

- Temos informações de altura de homens e mulheres, dados fictícios, para explicar para alguém de outra cultura e país, como somos fisicamente.

Considere que os dados estão divididos em:
Altura, em centímetros (valores numéricos)

1. Smoothed Density pode ser usado para comparar distribuições mais facilmente



Modelagem e Análise de Dados

Distribuição dos Dados

A distribuição Normal

Sabemos que geralmente a média e o desvio padrão fazem um bom papel de resumo dos dados. Isso é verdade graças à distribuição normal.

Vários fenômenos aproximam a distribuição de uma Normal:

- Pressão sanguínea
- Pesos
- Alturas
- Notas de avaliações padronizadas
- Erros de medidas experimentais
- E muitos outros...

Variáveis Aleatórias!

Modelagem e Análise de Dados

Distribuição dos Dados

A distribuição Normal

- A distribuição é simétrica, centrada na média
- A maioria dos valores (~95%) estão até 2 desvios padrão a partir da média
- A imagem ao lado tem média 0 e desvio padrão 1.

As definições nos implicam que: Se um conjunto de dados é aproximado por uma distribuição normal, todas as informações que precisamos para descrever a distribuição pode ser focada em: média e desvio padrão.

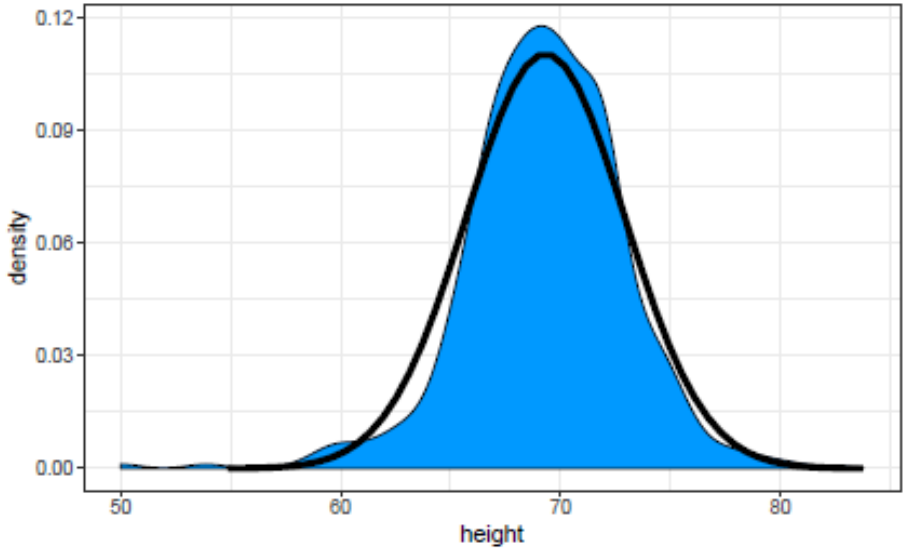
Modelagem e Análise de Dados

Distribuição dos Dados

A distribuição Normal

Considerando o mesmo exemplo das alturas, podemos observar que a densidade apresentada pode ser muito bem representada por uma distribuição Normal.

Neste caso, média = 69,3 e desvio padrão = 3,6



Modelagem e Análise de Dados

Distribuição dos Dados

A distribuição Normal

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi}s} e^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2} dx$$

→ Unidades padronizadas

Para dados que são aproximadamente distribuídos segundo uma Normal é conveniente que estejam em termos de unidades padronizadas.

$$z = \frac{x - m}{s}$$

Assim, não importa a unidade original dos dados.
As regras serão sempre válidas: média em $z = 0$, um dos maiores em $z \approx 2$, um dos menores em $z \approx -2$ e os que ocorrem raramente em $z > 3$ e $z < -3$. **Obs:** A função StandardScaler do scikitLearn faz isso!

Modelagem e Análise de Dados

Média: valor médio do conjunto. É a soma de todos os valores e divide pelo número total de observações.

Médiana: é o valor que separa um conjunto de dados em duas metades iguais, 50% abaixo e 50% acima deste valor.

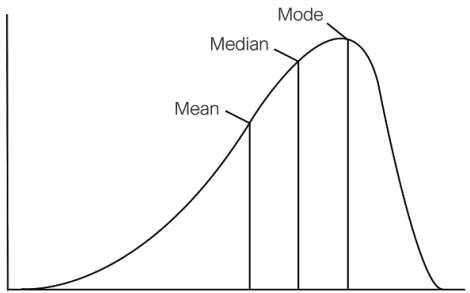
Moda: é o valor que ocorre com maior frequência em um conjunto de dados.

Distribuição dos Dados

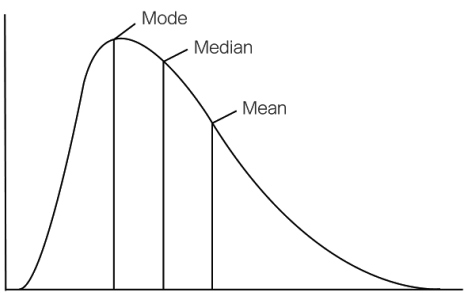
A distribuição Skewed

Os dados podem ter uma distribuição não “centrada”. A maioria dos valores estarão perto do início ou do fim da distribuição.

1. Left-Skewed
2. Right-Skewed



Left-Skewed (Negative Skewness)



Right-Skewed (Negative Skewness)

Modelagem e Análise de Dados

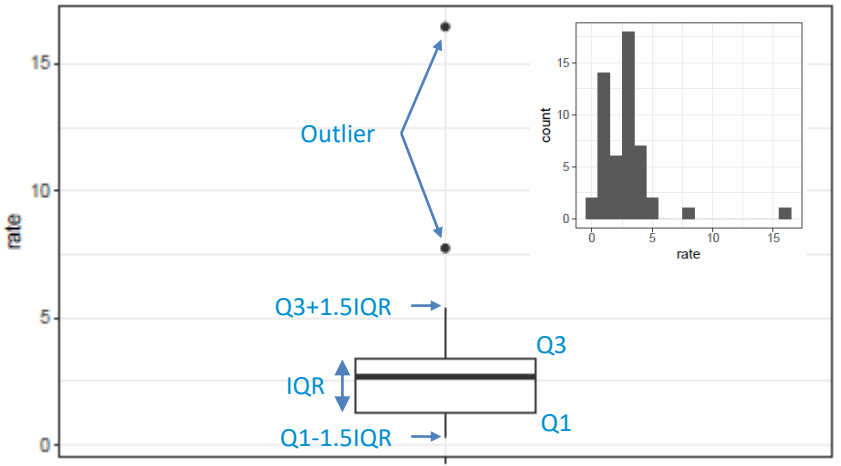
Distribuição dos Dados

Indo além do resumo dos dois números e identificando Outliers

Para os casos em que a distribuição Normal não representa bem os dados, principalmente, mais informações podem ser solicitadas, além da média e desvio padrão.

O BoxPlot foi introduzido para este cenário.

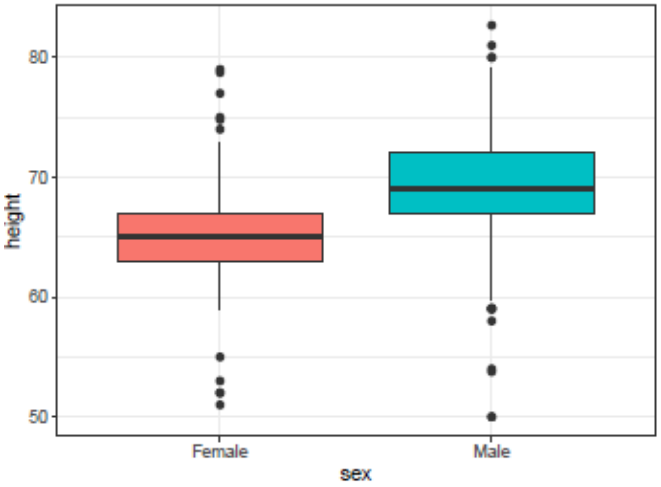
Cinco pontos de interesse:
- 25%, 50%, 75%, mínimo e máximo.
- Outliers também são apresentados, mas, podem ser ignorados para uma melhor visualização do BoxPlot.

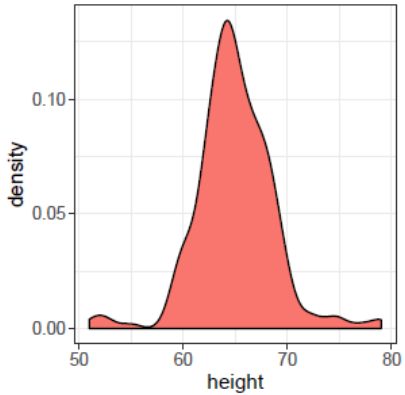


Modelagem e Análise de Dados

Distribuição dos Dados

Finalizando o exemplo...





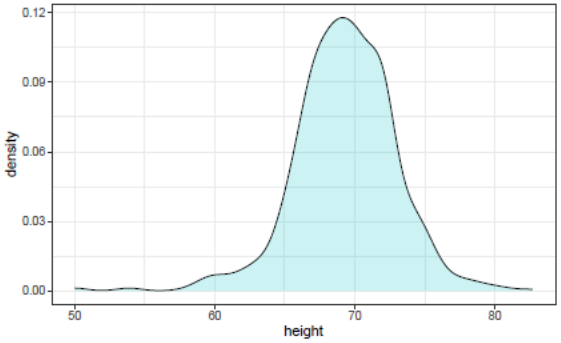
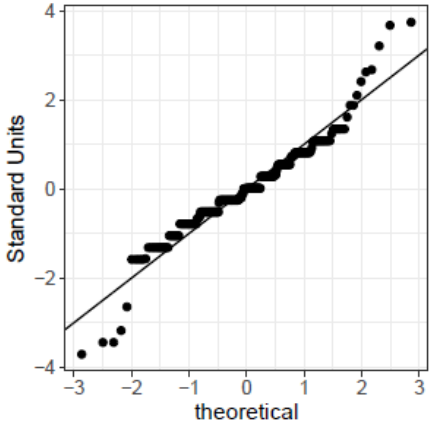


Gráfico Q-Q
Compara-se os quantis dos dados observados com os quantis da distribuição teórica (normal)



Análise de Dados

- Modelagem e Análise de Dados
- Estatística aplicada a análise de dados

Modelagem e Análise de Dados

EDA – Exploratory Data Analysis

Algumas análises iniciais:

pd.head() – ter uma amostra dos dados, entender os valores e nomes dos atributos
pd.describe() – descrição estatística sobre os dados: média, desvio padrão, mínimo, máximo, 25%, 50% e 75%
pd.boxplot() – apresenta os dados do describe() no formato gráfico
pd.quantile([0.0, 0.0]) – apresentar outros valores de quantís
pd.median() – calcula a mediana
pd.mean() – calcula a média
pd.std() – calcula o desvio padrão

Modelagem e Análise de Dados

EDA – Exploratory Data Analysis

Algumas análises iniciais:

Para dados categóricos
pd.unique() – retorna os valores possíveis para este atributo

Relação entre variáveis
plt.scatter(A, B)

Distribuição de uma variável
plt.hist(A, bins=X) – bins inicialmente pode ser $\sqrt{instâncias}$

Modelagem e Análise de Dados

EDA – Exploratory Data Analysis

Algumas análises iniciais:

Correlação
np.corrcoef(data) – matriz de coeficientes de correlação
plt.matshow(coefCorr)

Correlação não implica em Causalidade

O Check List da Análise de Dados

1. Respondendo as perguntas
    1. Você especificou o tipo de questão analítica do dados, antes de tocar no dados?
    2. Você especificou as métricas de sucesso (ou erro) no início?
    3. Você entendeu o contexto das questões e a aplicação científica ou de negócio?
    4. Você registrou o modelo?
    5. Você considerou se as questões podem ser respondidas com os dados disponíveis?

2. Checando os dados
    1. Você plotou resumos dos dados, uni ou multivariados?
    2. Você verificou os outliers?
    3. Você identificou dados faltantes?

O Check List da Análise de Dados

1. Organizando os dados
    1. Cada variável é uma coluna?
    2. Cada instância (ou observação) é uma linha?
    3. Diferentes tipos de dados aparecem em diferentes tabelas?
    4. Todos os parâmetros, unidades e funções aplicadas aos dados foram registrados?

2. Análise exploratória
    1. Você identificou os valores faltantes?
    2. Fez gráficos (histogramas, densidades, boxplots, etc)
    3. Considerou a correlação?
    4. Verificou as unidades e faixa de valores?
    5. Considerou fazer o gráfico em escala logarítmica?

O Check List da Análise de Dados

1. Inferência
    1. Você identificou o quão grande a população que você quer entender é?
    2. Você avaliou se as amostras são representativas?

2. Predição
    1. Definiu de antemão os erros?
    2. Dividiu os dados em treinamento e teste?
    3. Usou validação cruzada, agregação ou reamostragem nos dados de treinamento?
    4. Você criou novas *features*?
    5. Você estimou parâmetros? Estes estão fixos para validação?
    6. O modelo final foi o usado para validação e relatório do resultado?

O Check List da Análise de Dados

1. Análise
    1. Descreva a questão de interesse
    2. Descreva o *dataset*, o desenho do experimento e as questões a serem respondidas
    3. Descreva o tipo de análise frente às perguntas
    4. Descreva o modelo exato que irá ajustar
    5. Inclua análises com incertezas

2. Apresentação dos resultados
    1. Imagens valem mais que mil palavras
    2. Apresentação

inatel

inateloficial

ascominatel

inatel.tecnologias

company/inatel

Inatel

Inatel - Instituto Nacional de Telecomunicações
Campus em Santa Rita do Sapucaí - MG - Brasil
Av. João de Camargo, 510 - Centro - 37540-000
+55 (35) 3471 9200

Escritório em São Paulo - SP - Brasil
WTC Tower, 18º andar - Conjunto 1811/1812
Av. das Nações Unidas, 12.551 - Brooklin Novo - 04578-903
+55 (11) 3043 6015