

Análise Exploratória de Dados

Formulação de perguntas de dados preliminares

Felipe Figueiredo

Instituto Nacional de Traumatologia e Ortopedia

Sumário

1 Análise Exploratória

- EDA
- Tabelas
- Gráfico de Dispersão
- Histogramas
- Boxplot
- Referências

Paradigmas de Análises de Dados

Estudos quantitativos exigem coleta e análise de dados

- EDA – Análise Exploratória de Dados
- CDA – Análise Confirmatória de Dados

Análise Exploratória de Dados

- Formalizado por John W. Tukey nos anos 1970
- Objetivo: formular perguntas com base nos dados disponíveis
- Perguntas que podem ser respondidas pela análise dos dados

O que é

EDA é uma filosofia/approach para

- maximizar a compreensão/insight sobre um dataset
- descobrir estruturas
- identificar variáveis importantes
- detectar outliers e anomalias

Fonte: NIST Handbook (1998)

Tukey, 1980

"Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. (...) Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, (...) NOT a bundle of techniques (...)."

Tukey (1980) - We Need Both Exploratory and Confirmatory

Para Tukey:

- Não basta fazer a análise confirmatória, nem a exploratória: **ambas** são necessárias
- Paradigma: pergunta → resposta é inadequado
- "Encontrar a pergunta é por vezes mais importante que encontrar a resposta"

(*) question → design → collection →
analysis → answer

- 1 Como as perguntas são geradas?
- 2 Como os desenhos (experimentais) são guiados?
- 3 Como a coleta de dados é monitorada?

Questões sobre o paradigma confirmatório



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

Respostas: Geralmente por ...

- ① *insights* teóricos e a exploração de dados anteriores (e.g., pesquisa bibliográfica)
- ② informação qualitativa disponível obtida da exploração de dados anteriores
- ③ exploração dos dados, conforme são obtidos, buscando comportamento “inesperado”

Explorar...



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

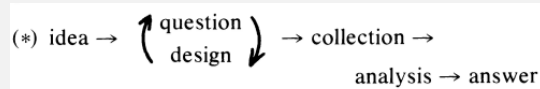
- A chave então é explorar os dados
- Explorar antes, durante e depois da análise confirmatória
- Busca de pistas, idéias e eventualmente conclusões preliminares (*hipóteses!*)

Paradigma alternativo



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências



Tukey sugere que:

- Antes de termos uma pergunta, temos uma idéia (a ser formalizada)
- Pergunta formal depende dos dados disponíveis
- Questão pragmática, independe do desejo ou vontade

Exemplo



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

Example

- Idéia: uma certa droga ajuda em uma doença
 - Queremos testar/confirmar isso...
 - ... com consistência estatística na resposta
- Idéia preliminar informal, vaga
 - Geralmente em termos de linguagem coloquial
 - Não pode ser avaliada com suporte estatístico

Exemplo



Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

Desejo: pergunta geral, de amplo espectro e implicações profundas

Example

“Dos pacientes que morreriam em até três anos desta doença, que fração poderia ser salva por este tratamento?”

- Dificuldade técnica. . .
- . . . nenhum design pode isolar essas pessoas para um experimento

Exemplo



Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

O que **pode** ser perguntado está limitado por:

- Idade e sexo dos pacientes
- conjunto mínimo de sintomas
- ausência de outras condições potencialmente fatais
- tipos de pacientes que podem ser encontrados/observados
- etc.

Formulação da pergunta



Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

- o que pode concretamente ser perguntado
- que desenhos são viáveis
- chance de um certo design resultar em resposta útil
- “Como eu estudo o que está acontecendo aqui?”

Por onde começar?



Análise
Exploratória
de Dados

Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

- tabelas
- gráficos dos dados brutos
- estatísticas descritivas simples
- procurar padrões

Example

Pacientes que tem uma efermidade grave, podem ser submetidos a um tratamento cirúrgico.

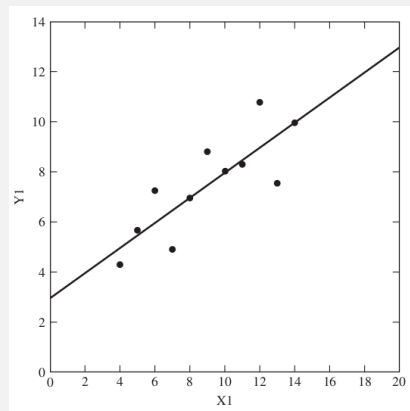
	Óbito	não óbito	Total
Cirurgia	3	1	4
não cirurgia	2	5	7
Total	5	6	11

Exercício

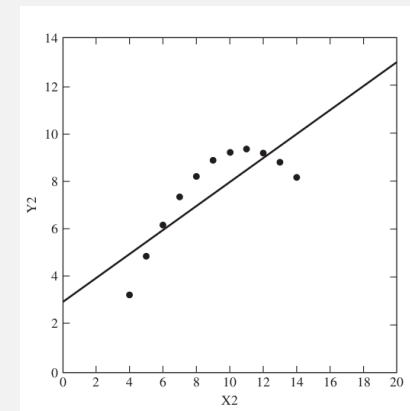
Formule uma pergunta sobre este contexto.

Gráficos de dispersão (ou scatter-plots):

- visualizar os dados pontuais diretamente
- identificar possíveis padrões ou tendências
- identificar visualmente possíveis outliers
- desenhar possíveis relações (modelos) sobre os dados



Fonte: Behrens, Yu (2003)



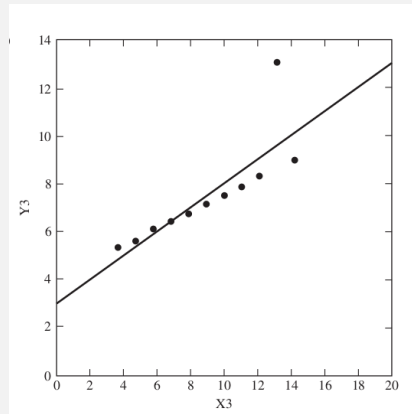
Fonte: Behrens, Yu (2003)

Anomalias



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências



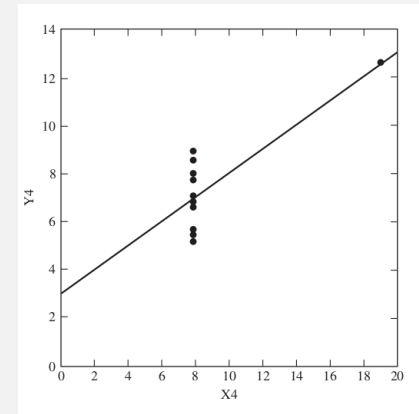
Fonte: Behrens, Yu (2003)

??



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências



Fonte: Behrens, Yu (2003)

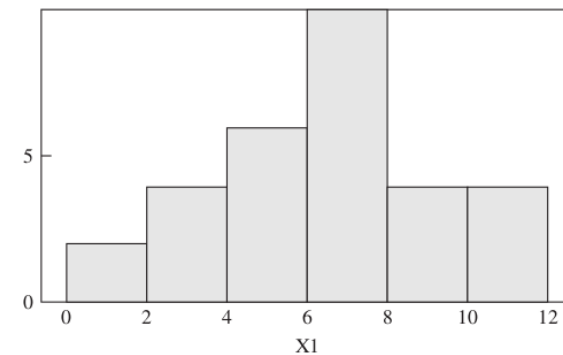
Histograma



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

- Gráfico de barras com frequências dos dados
- visualização prática da distribuição dos dados
- identificar simetria, tendência central, dispersão, etc



Fonte: Behrens, Yu (2003)



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

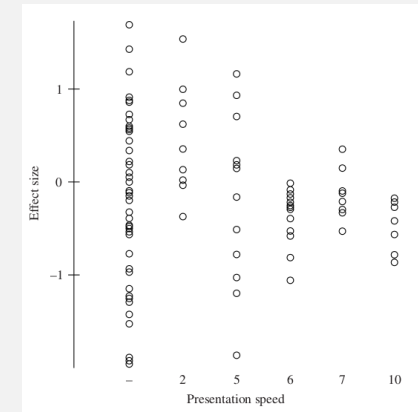
Boxplot



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

- caixa que contém 50% dos dados...
- ... e segmentos verticais que englobam a maior parte dos dados
- elementos fora deste intervalo devem ser investigados como possíveis outliers
- Ideal para grandes quantidades de dados



Fonte: Behrens, Yu (2003)



Análise
Exploratória
de Dados
Felipe
Figueiredo

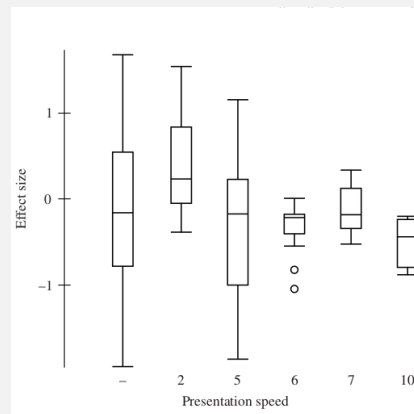
Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

O boxplot



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências



Fonte: Behrens, Yu (2003)

Referências



Análise
Exploratória
de Dados
Felipe
Figueiredo

Análise
Exploratória
EDA
Tabelas
Gráfico de Dispersão
Histogramas
Boxplot
Referências

- NIST Handbook (1998), Exploratory Data Analysis, cap 1 - <http://www.itl.nist.gov/div898/handbook/eda/section1/eda1.htm> (Acessado em 10/09/2015)
- Tukey (1980), We need both exploratory and confirmatory, <http://www-ece.rice.edu/~fk1/classes/ELEC697/TukeyEDA.pdf> (Acessado em 10/09/2015)
- Behrens, Yu (2003), Exploratory Data Analysis, cap 2 - Research Methods in Psychology