# Análise Exploratória de Dados

## Formulação de perguntas de dados preliminares

Felipe Figueiredo

Instituto Nacional de Traumatologia e Ortopedia

# Sumário

# Sumário

# Paradigmas de Análises de Dados

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

# Paradigmas de Análises 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

# 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

# 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

# Análise Exploratória de 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)

# Análise Exploratória de 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)

# Análise Exploratória de 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)

# Análise Exploratória de 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)

# Análise Exploratória de 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)

# Análise Exploratória de Dados



**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 atitude, (. . . ) NOT a bundle of techniques (. . . )."*

Tukey (1980) - We Need Both Exploratory and Confirmatory

# Análise Exploratória de Dados



Para Tukey:

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

# Análise Exploratória de Dados



Para Tukey:

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

# Análise Exploratória de Dados



Para Tukey:

- Não basta fazer a análise confirmatória, nem a exploratória: ambas são necessárias
- Paradigma: pergunta $\rightarrow$ resposta é inadequado
- “Encontrar a pergunta é por vezes mais importante que encontrar a resposta”

# Paradigma linear

(*) 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?

# Paradigma linear

(*) 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?

# Paradigma linear

(*) 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



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”

# Questões sobre o paradigma confirmatório

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”

# Questões sobre o paradigma confirmatório

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...

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

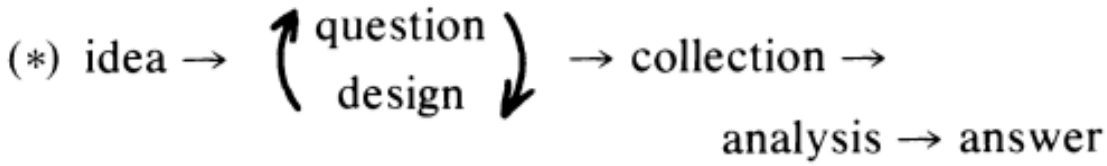
# Explorar...

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

# Explorar...

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

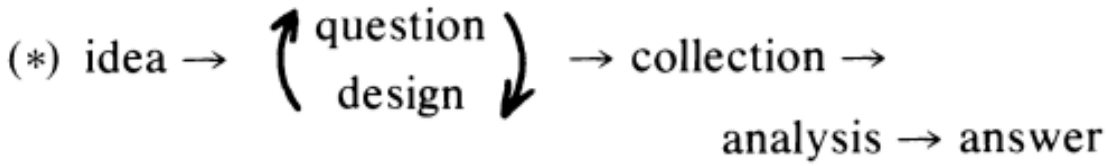
# Paradigma alternativo




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

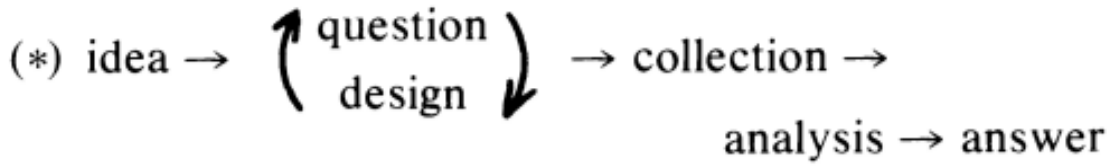
# Paradigma alternativo




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

# Paradigma alternativo




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



**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

## 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



**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

**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

**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

**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

**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

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

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

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

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.

# Exemplo

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.

# Exemplo

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.

# Exemplo

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.

# Exemplo

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

- 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?”

# Formulação da pergunta

- 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?”

# Formulação da pergunta

- 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?”

# Formulação da pergunta

- 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?

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

# Por onde começar?

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

# Por onde começar?

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

# Sumário

# Gráficos de dispersão



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

# Gráficos de dispersão



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

# Gráficos de dispersão



Gráficos de dispersão (ou scatter-plots):

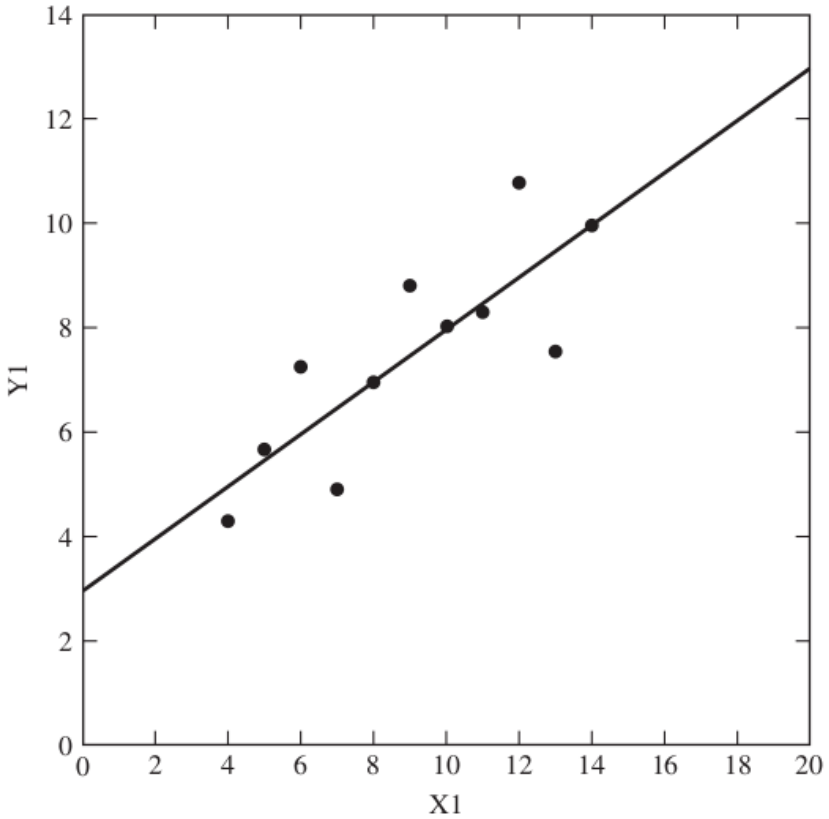
- visualizar os dados pontuais diretamente
- identificar possíveis padrões ou tendências
- identificar visualmente possíveis outliers
- desenhar possíveis relações (modelos) sobre os dados

# Gráficos de dispersão

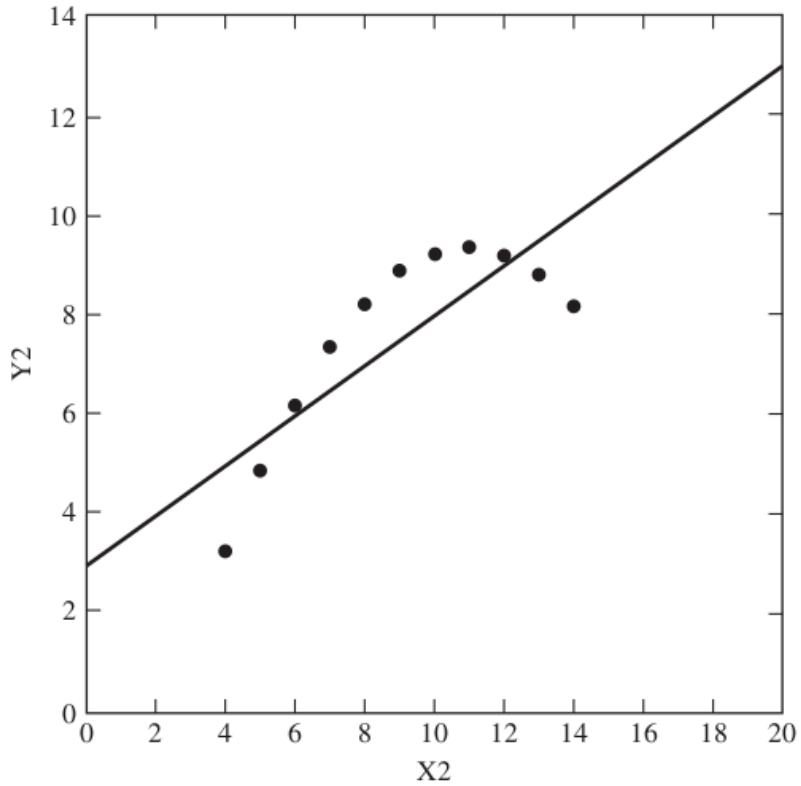


Gráficos de dispersão (ou scatter-plots):

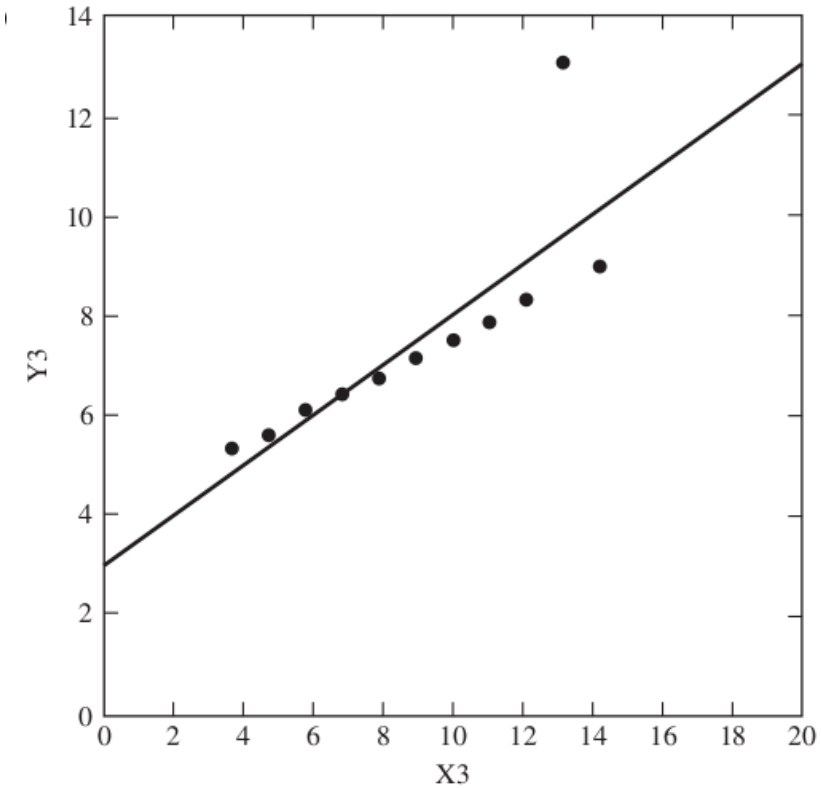
- visualizar os dados pontuais diretamente
- identificar possíveis padrões ou tendências
- identificar visualmente possíveis outliers
- desenhar possíveis relações (modelos) sobre os dados



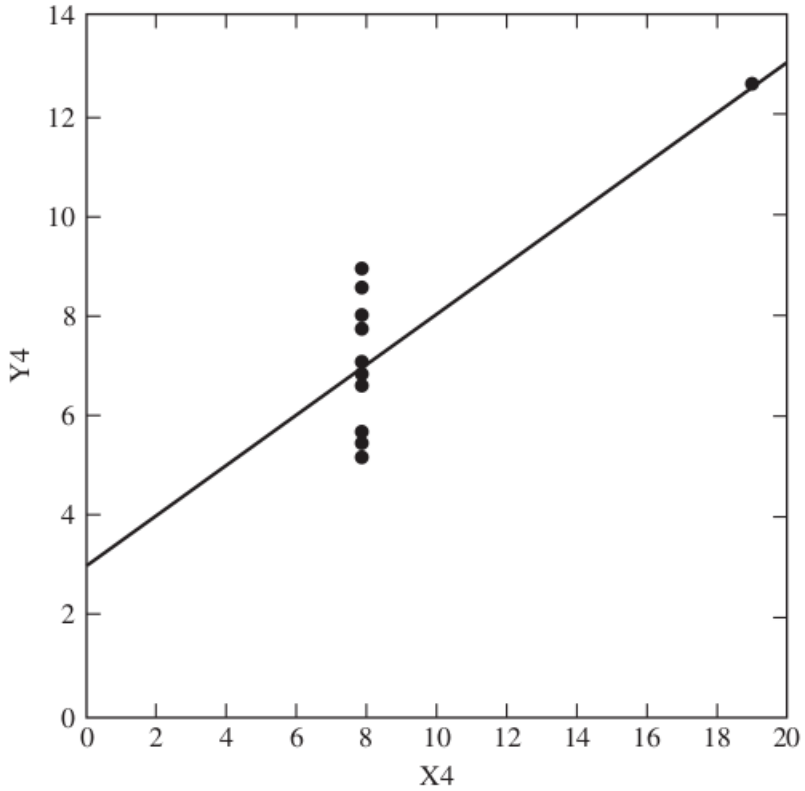

Fonte: Behrens, Yu (2003)




Fonte: Behrens, Yu (2003)




Fonte: Behrens, Yu (2003)




Fonte: Behrens, Yu (2003)

# Sumário

# Histograma

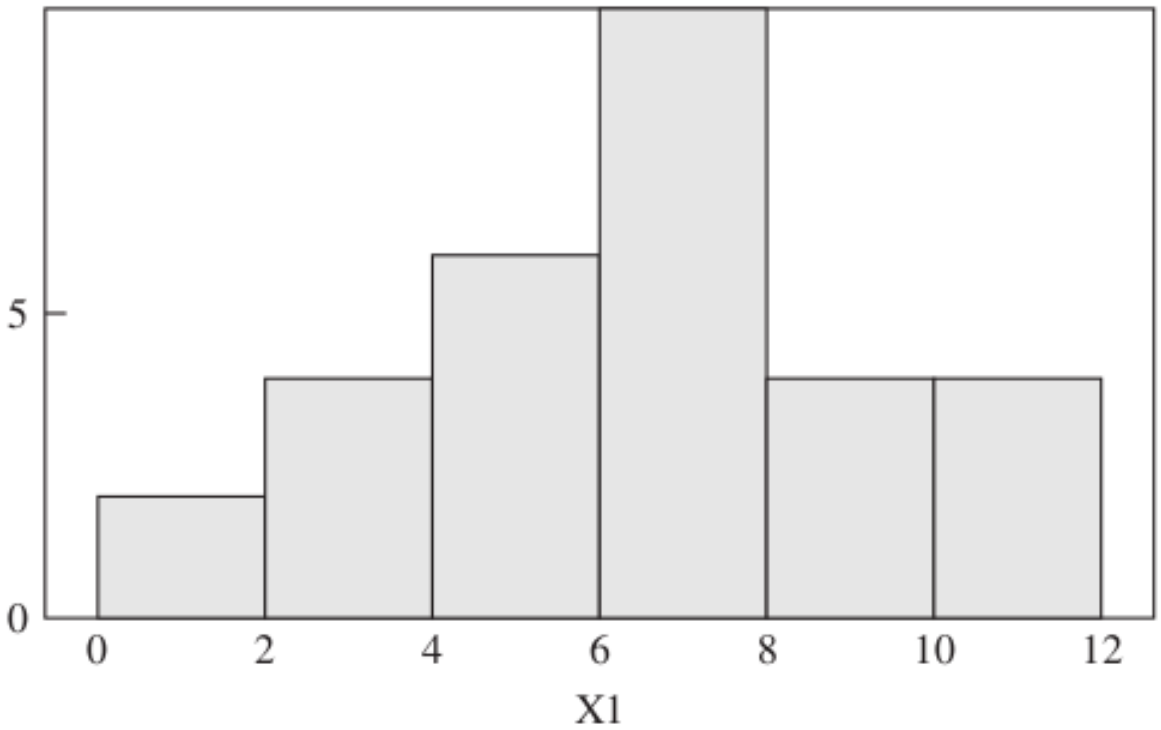
- Gráfico de barras com frequências dos dados
- visualização prática da distribuição dos dados
- identificar simetria, tendência central, dispersão, etc

# Histograma

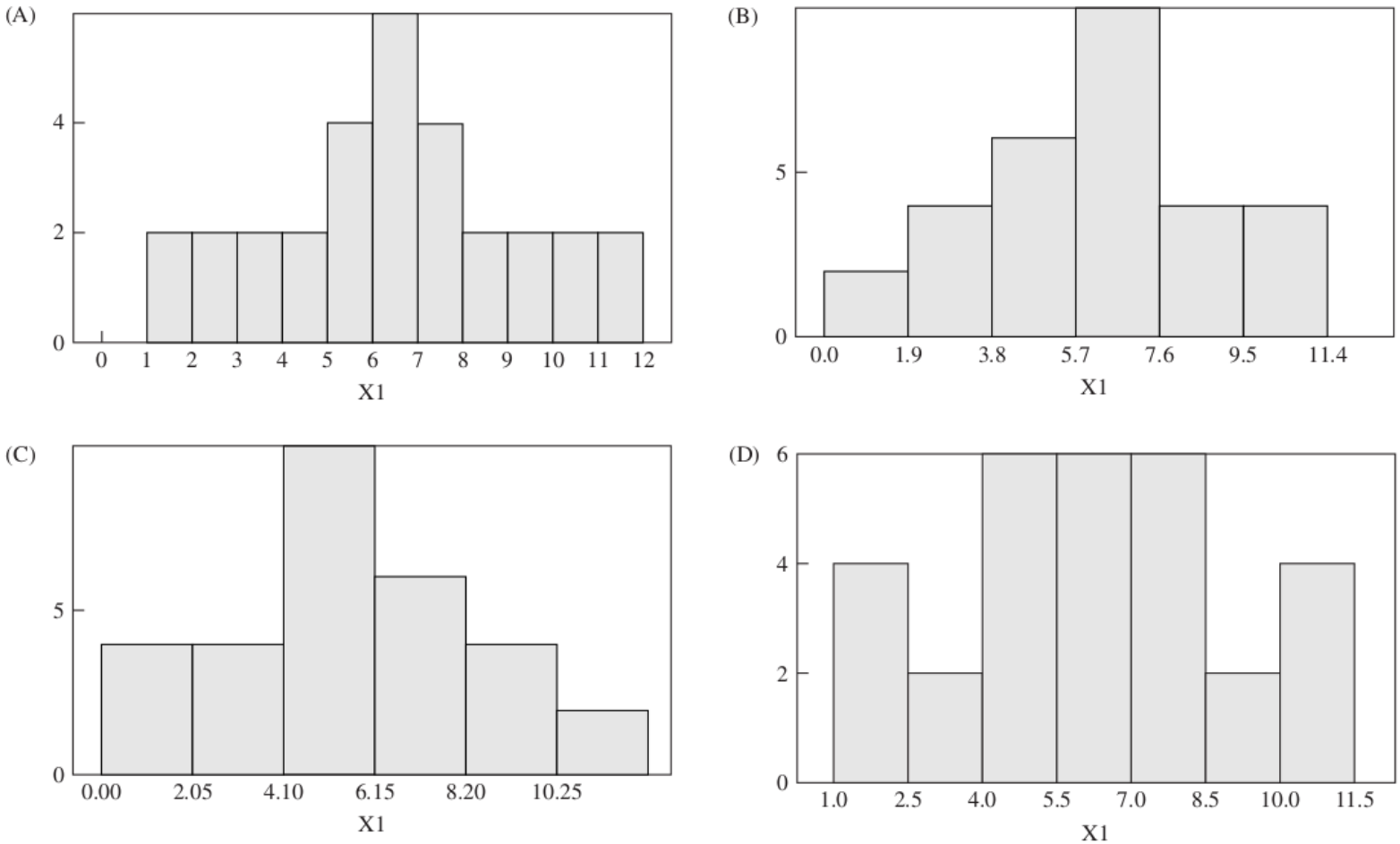
- Gráfico de barras com frequências dos dados
- visualização prática da distribuição dos dados
- identificar simetria, tendência central, dispersão, etc

# Histograma

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




Fonte: Behrens, Yu (2003)




Fonte: Behrens, Yu (2003)

# Sumário

# Boxplot

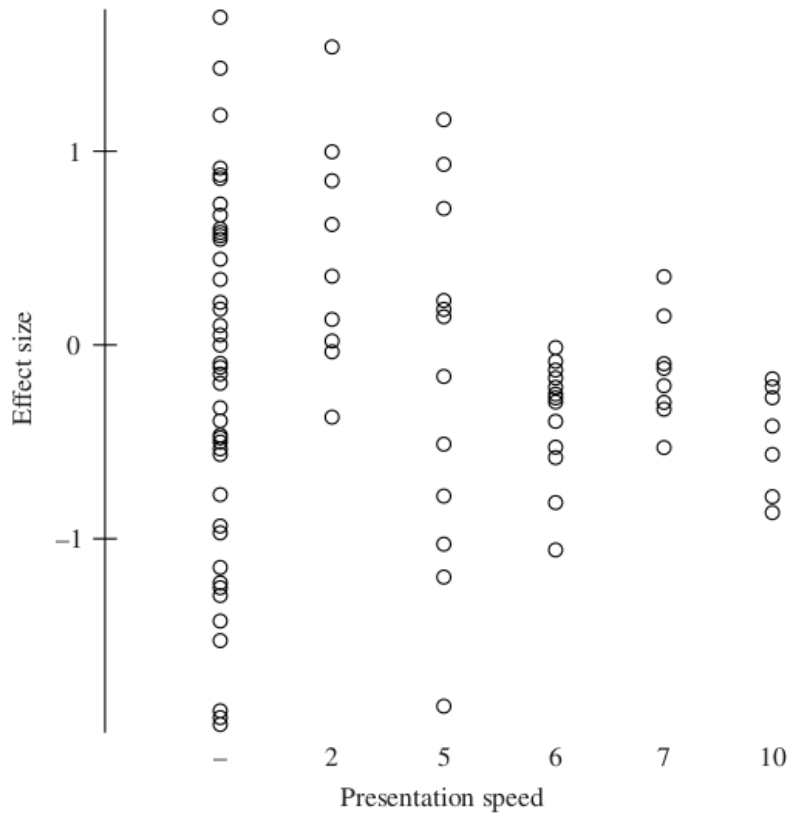
- caixa que contém 50% dos dados...
- ...e segmentos verticais que englobam a maior parte dos dados
- elementos fora deste intervalo devem ser investigados como possíveis outliers

# Boxplot

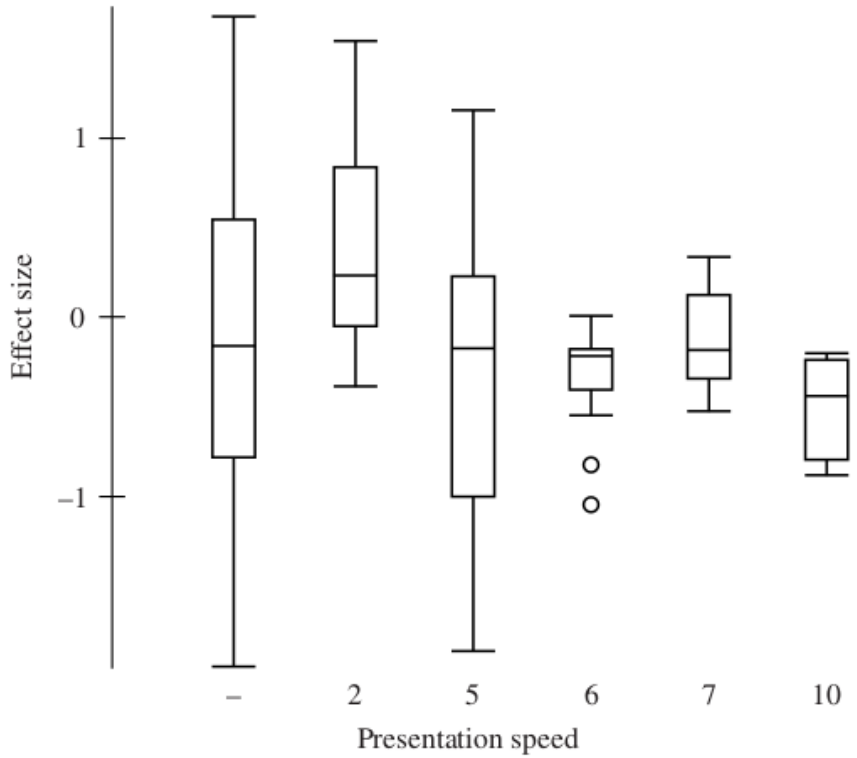
- caixa que contém 50% dos dados...
- ...e segmentos verticais que englobam a maior parte dos dados
- elementos fora deste intervalo devem ser investigados como possíveis outliers

# Boxplot

- 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




Fonte: Behrens, Yu (2003)




Fonte: Behrens, Yu (2003)

# Sumário

# 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