

Tópicos em resultados preliminares

Análise Exploratória de Dados

Felipe Figueiredo

Instituto Nacional de Traumatologia e Ortopedia

1 Análise Exploratória

- EDA
- Tabelas
- Figuras
- Exercício
- Resumo
- Referências

2 Aprofundamento

- Aprofundamento

Discussão da aula passada



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

Aprofundamento

Discussão da leitura obrigatória da aula passada

1 Análise Exploratória

- EDA
- Tabelas
- Figuras
- Exercício
- Resumo
- Referências

2 Aprofundamento

- Aprofundamento

Evidências

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

Sherlock Holmes

Estudos quantitativos requerem coleta e análise de dados

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

Estudos quantitativos requerem coleta e análise de dados

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

- 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

Uma filosofia/approach para

- insight sobre um dataset
- descobrir estruturas/padrões
- identificar variáveis importantes
- detectar outliers e anomalias

NIST Handbook (1998)

Do resumo...

"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

Do resumo...

"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

- ① Um paradigma incompleto
- ② Origem das ideias
- ③ Perguntas importantes
- ④ A investigação abrangente
- ⑤ Uma máxima
- ⑥ Análise confirmatória

Tukey, 1980

(*) question \rightarrow design \rightarrow collection \rightarrow
analysis \rightarrow answer

Tópicos em resultados preliminares

Felipe
Figueiredo

Análise Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

Este paradigma simplista presume que...

- Sabemos a pergunta “correta” no início
- Ignora questões importantes sobre o processo investigativo
 - Como as perguntas são geradas?
 - Como os desenhos (experimentais) são guiados?
 - Como a coleta de dados é monitorada?

Tukey, 1980

Paradigma linear – incompleto



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

(*) question → design → collection →

analysis → answer

Este paradigma simplista presume que...

- Sabemos a pergunta “correta” no início
- Ignora questões importantes sobre o processo investigativo
 - Como as perguntas são geradas?
 - Como os desenhos (experimentais) são guiados?
 - Como a coleta de dados é monitorada?

Tukey, 1980

Paradigma linear – incompleto



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

(*) question → design → collection →

analysis → answer

Como as perguntas são geradas?

Geralmente por *insights* teóricos e a exploração de dados anteriores (e.g., pesquisa bibliográfica)

Tukey, 1980

Paradigma linear – incompleto



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

(*) question \rightarrow design \rightarrow collection \rightarrow

analysis \rightarrow answer

Como os desenhos (experimentais) são guiados?

Geralmente por informação qualitativa disponível obtida da exploração de dados anteriores

Tukey, 1980

Paradigma linear – incompleto



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

(*) question → design → collection →

analysis → answer

Como a coleta de dados é monitorada?

Geralmente pela exploração dos dados, conforme são obtidos, buscando comportamento “inesperado”

Tukey, 1980

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

Tukey, 1980

A origem das ideias



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

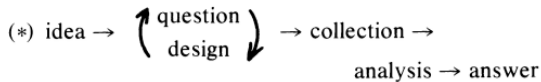
Figuras

Exercício

Resumo

Referências

Aprofundamento



Tukey sugere que:

- Antes de termos uma pergunta, temos uma ideia (a ser formalizada)¹
- Pergunta formal depende dos dados disponíveis
- Questão pragmática, independe do desejo ou vontade

Tukey, 1980

¹ Assim como sua “proposta de pergunta 1/2”

A origem das ideias



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

Exemplo

- Ideia: uma certa droga ajuda em uma doença
- Queremos testar/confirmar isso...
- ...com consistência estatística na resposta

- Ideia preliminar informal, vaga
- Geralmente em termos de linguagem coloquial
- Não pode ser avaliada com suporte estatístico

Tukey, 1980

A origem das ideias



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

Exemplo

- Ideia: uma certa droga ajuda em uma doença
- Queremos testar/confirmar isso...
- ... com consistência estatística na resposta

- Ideia preliminar informal, vaga
- Geralmente em termos de linguagem coloquial
- Não pode ser avaliada com suporte estatístico

Tukey, 1980

A origem das ideias



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

Exemplo

- Ideia: uma certa droga ajuda em uma doença
- Queremos testar/confirmar isso...
- ... com consistência estatística na resposta

- Ideia preliminar informal, vaga
- Geralmente em termos de linguagem coloquial
- Não pode ser avaliada com suporte estatístico

Tukey, 1980

A origem das ideias



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

Desejo: pergunta geral, de amplo espectro e implicações profundas

Exemplo

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

- Dificuldade técnica²...
- ...nenhum design pode isolar essas pessoas para um experimento

Tukey, 1980

²Neste exemplo, questão ética

A origem das ideias



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

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.

Tukey, 1980

A origem das ideias



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

- 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?”

Tukey, 1980

Por onde começar?



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

Aprofundamento

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

1 Análise Exploratória

- EDA
- **Tabelas**
- Figuras
- Exercício
- Resumo
- Referências

2 Aprofundamento

- Aprofundamento

Exemplo

Table 2. Patient Clinical Characteristics

Characteristics	Median Sternotomy (n = 84)	Right Minithoracotomy (n = 119)	p Value
Patient characteristics:			
Age, years (median, IQR)	80 (78–84)	79 (77–83)	0.12
Males (%)	37 (44)	47 (39)	0.18
Body mass index (IQR)	26.2 (23.9–29.2)	26.5 (23.1–29.7)	0.95
Preoperative creatinine (IQR)	1.02 (0.87–1.3)	1.02 (0.86–1.25)	0.65
Ejection fraction (median, IQR)	0.55 (0.46–0.60)	0.58 (0.50–0.63)	0.29
Diabetes mellitus (%)	20 (23.8)	32 (26.9)	0.31
Hypertension (%)	80 (95.2)	109 (91.6)	0.62
Peripheral vascular disease (%)	8 (9.5)	7 (5.9)	0.33
Cerebrovascular disease (%)	9 (10.7)	19 (16)	0.29
Prior coronary bypass graft surgery (%)	10 (11.9)	12 (10.1)	0.68
Prior valve surgery (%)	8 (9.5)	8 (6.7)	0.47
Prior heart failure (%)	47 (56)	43 (36.1)	0.005
Procedural characteristics:			
Mitral valve surgery	49%	51%	0.75
Aortic valve surgery	51%	49%	0.75
Cardiopulmonary bypass time minutes (IQR)	86 (39–268)	118 (67–186)	<0.001
Cross-clamp time minutes (IQR)	61 (25–156)	84 (40–154)	<0.001

IQR = interquartile range.

Exemplo

Pacientes que tem uma enfermidade 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.

1 Análise Exploratória

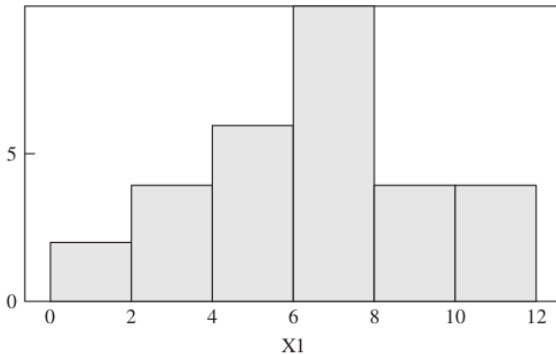
- EDA
- Tabelas
- **Figuras**
- Exercício
- Resumo
- Referências

2 Aprofundamento

- Aprofundamento

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

Exemplo



Behrens, Yu (2003)

Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

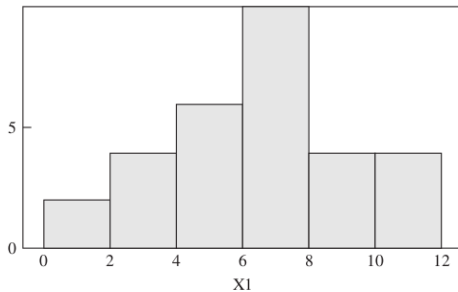
Exercício

Resumo

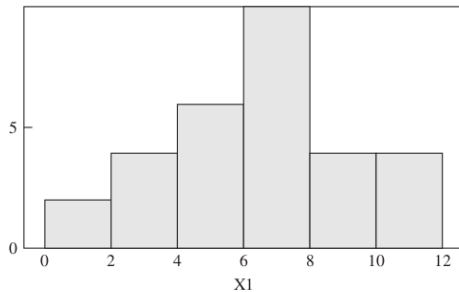
Referências

Aprofundamento

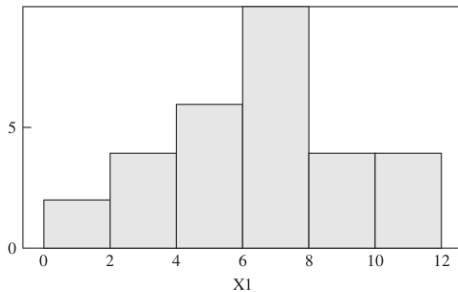
- Mensurações mais frequentes no centro
- Mensurações altas/baixas menos frequentes...
- ... com frequências semelhantes (simetria)
- Ideia da variabilidade das mensurações ("largura")



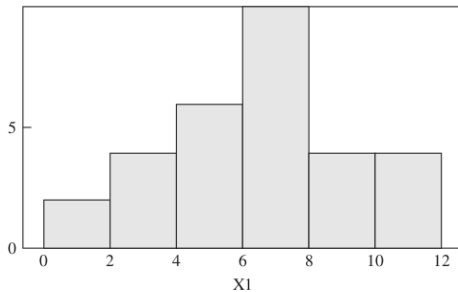
- **Mensurações mais frequentes no centro**
- Mensurações altas/baixas menos frequentes...
- ... com frequências semelhantes (simetria)
- Ideia da variabilidade das mensurações ("largura")



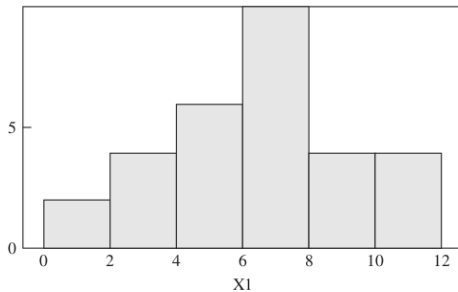
- Mensurações mais frequentes no centro
- Mensurações altas/baixas menos frequentes...
- ... com frequências semelhantes (simetria)
- Ideia da variabilidade das mensurações ("largura")



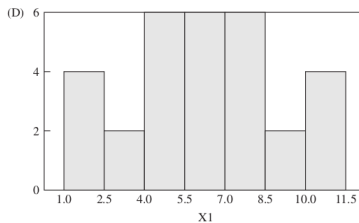
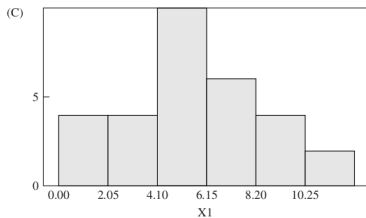
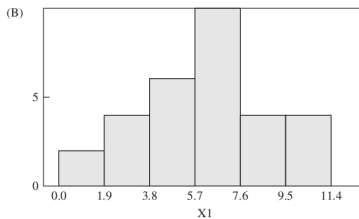
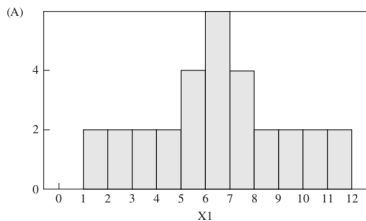
- Mensurações mais frequentes no centro
- Mensurações altas/baixas menos frequentes...
- ... com frequências semelhantes (simetria)
- Ideia da variabilidade das mensurações ("largura")



- Mensurações mais frequentes no centro
- Mensurações altas/baixas menos frequentes...
- ... com frequências semelhantes (simetria)
- **Ideia da variabilidade das mensurações ("largura")**



Distribuições de dados podem ter várias formas



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

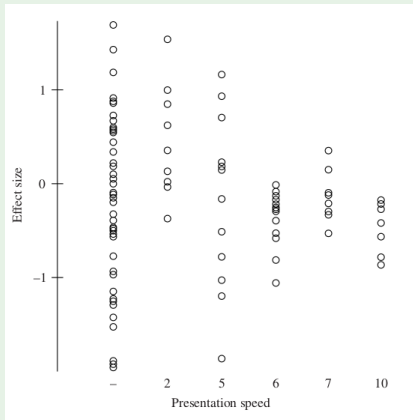
Aprofundamento

- Mensurações feitas em dois ou mais grupos³
- caixa que contém 50% dos dados...
- ... e segmentos verticais que englobam a maior parte dos dados
- mensurações fora dos limites \Rightarrow investigar possíveis outliers⁴
- Ideal para grandes quantidades de dados

³ e.g., exposição/tratamento 1, tratamento 2, controle...

⁴ erros de mensuração, imputação, viés de seleção/amostragem, observações raras...?

Exemplo



Behrens, Yu (2003)

Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

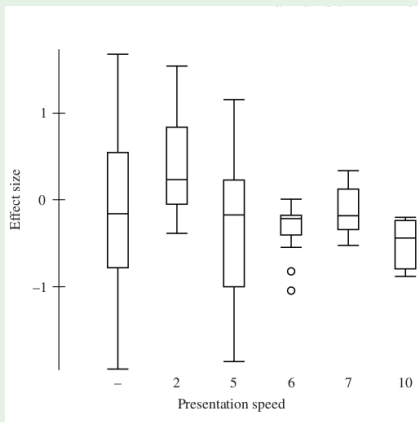
Exercício

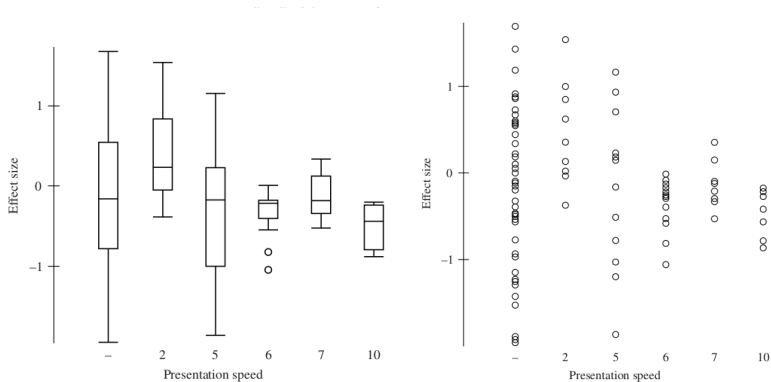
Resumo

Referências

Aprofundamento

Exemplo





Gráficos de dispersão



Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

Figuras

Exercício

Resumo

Referências

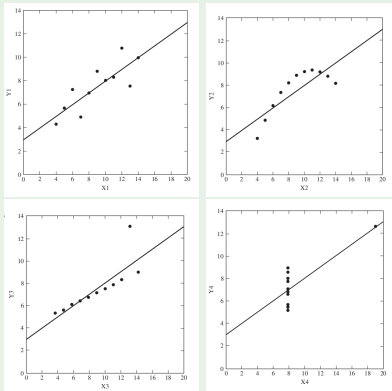
Aprofundamento

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

A álgebra mente...
... portanto figuras são necessárias

Behrens, Yu (2003)

Datasets didáticos de Anscombe, 1973



Quatro datasets com perfis
completamente diferentes \Rightarrow mesma reta de melhor ajuste

Behrens, Yu (2003)
NIST Handbook (1998)

Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

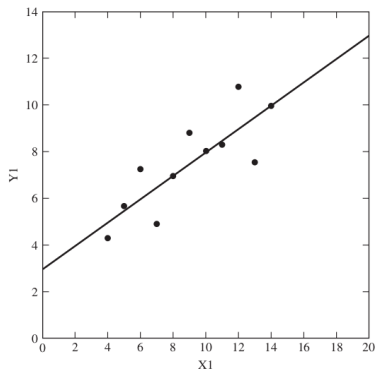
Figuras

Exercício

Resumo

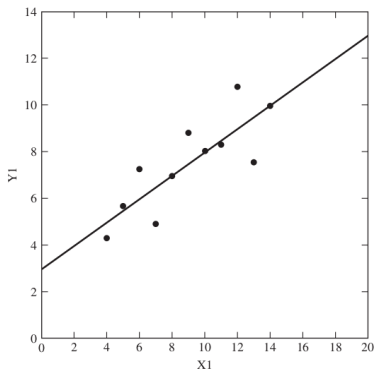
Referências

Aprofundamento

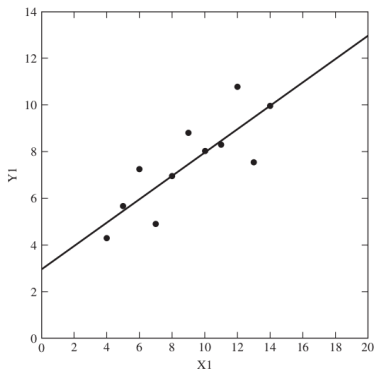


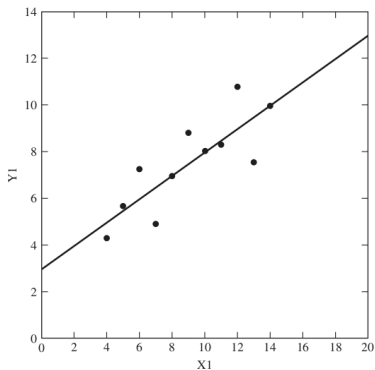
- Relação "claramente" linear...
- ... alguma dispersão
 - Não há justificativa para um modelo mais complexo (quadrático, etc...)
- Não há outliers
- Distância vertical à reta (Y) semelhante ao longo da faixa (X)
 - Não é necessário aplicar ponderações ou transformações

- Relação "claramente" linear...
- ... alguma dispersão
 - Não há justificativa para um modelo mais complexo (quadrático, etc...)
- Não há outliers
- Distância vertical à reta (Y) semelhante ao longo da faixa (X)
 - Não é necessário aplicar ponderações ou transformações

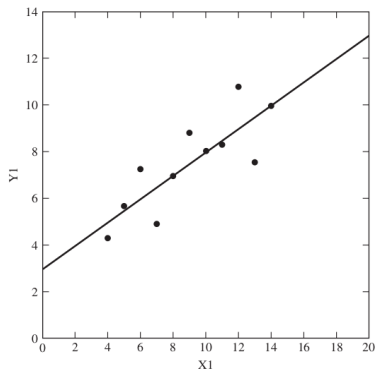


- Relação "claramente" linear...
- ... alguma dispersão
 - Não há justificativa para um modelo mais complexo (quadrático, etc...)
- Não há outliers
- Distância vertical à reta (Y) semelhante ao longo da faixa (X)
 - Não é necessário aplicar ponderações ou transformações

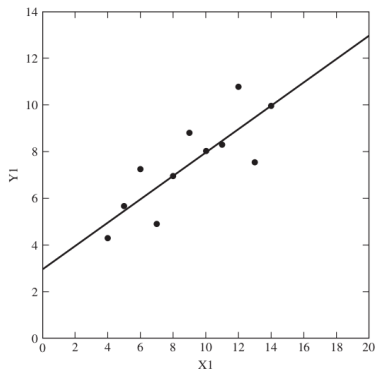




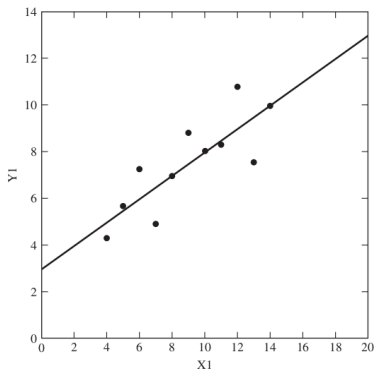
- Relação "claramente" linear...
- ... alguma dispersão
 - Não há justificativa para um modelo mais complexo (quadrático, etc...)
- Não há outliers
- Distância vertical à reta (Y) semelhante ao longo da faixa (X)
 - Não é necessário aplicar ponderações ou transformações



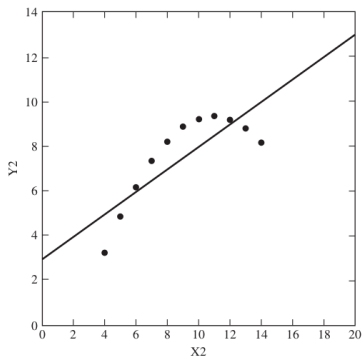
- Relação "claramente" linear...
- ... alguma dispersão
 - Não há justificativa para um modelo mais complexo (quadrático, etc...)
- **Não há outliers**
- Distância vertical à reta (Y) semelhante ao longo da faixa (X)
 - Não é necessário aplicar ponderações ou transformações



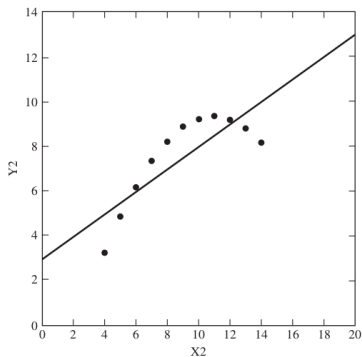
- Relação "claramente" linear...
- ... alguma dispersão
 - Não há justificativa para um modelo mais complexo (quadrático, etc...)
- Não há outliers
- Distância vertical à reta (Y) semelhante ao longo da faixa (X)
 - Não é necessário aplicar ponderações ou transformações



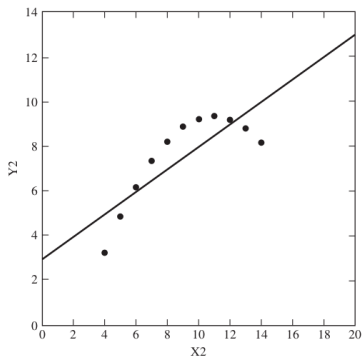
- Relação "claramente" linear...
- ... alguma dispersão
 - Não há justificativa para um modelo mais complexo (quadrático, etc...)
- Não há outliers
- Distância vertical à reta (Y) semelhante ao longo da faixa (X)
 - Não é necessário aplicar ponderações ou transformações



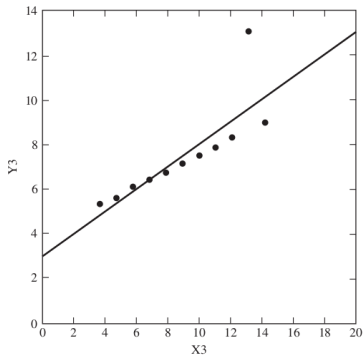
- Relação claramente não é linear
- Relação "claramente" quadrática



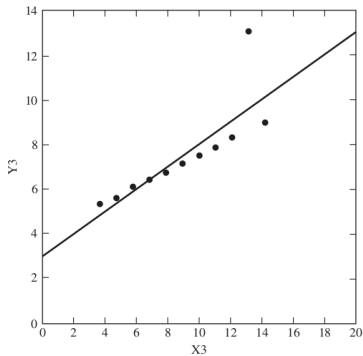
- Relação claramente não é linear
- Relação "claramente" quadrática



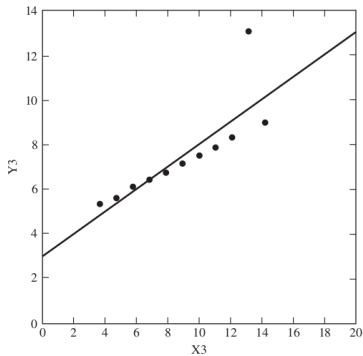
- Relação claramente não é linear
- Relação "claramente" quadrática



- "Claramente" possui um outlier
- ... que "puxa" a reta para cima



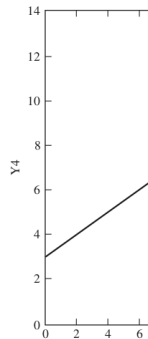
- “Claramente” possui um outlier
- ... que “puxa” a reta para cima



- "Claramente" possui um outlier
- ... que "puxa" a reta para cima

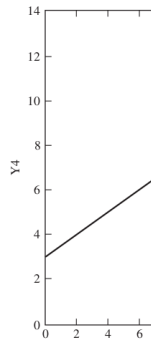
- CLARAMENTE vítima de um desenho experimental infeliz^a
- Um único ponto afastado do cluster de dados
- "cauda abanando o cachorro"

^aNão assistiu a aula de Planejamento/Protocolo



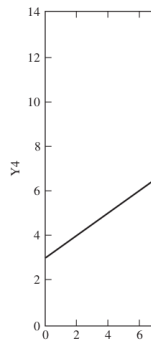
- CLARAMENTE vítima de um desenho experimental infeliz^a
- Um único ponto afastado do cluster de dados
- "cauda abanando o cachorro"

^aNão assistiu a aula de Planejamento/Protocolo



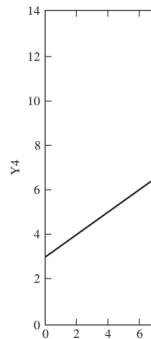
- CLARAMENTE vítima de um desenho experimental infeliz^a
- Um único ponto afastado do cluster de dados
- "cauda abanando o cachorro"

^aNão assistiu a aula de Planejamento/Protocolo



- CLARAMENTE vítima de um desenho experimental infeliz^a
- Um único ponto afastado do cluster de dados
- “cauda abanando o cachorro”

^aNão assistiu a aula de Planejamento/Protocolo



1 Análise Exploratória

- EDA
- Tabelas
- Figuras
- **Exercício**
- Resumo
- Referências

2 Aprofundamento

- Aprofundamento

REV BRAS ORTOP, 2013;48(5):441-447



ELSEVIER



Artigo Original

Enxerto ideal para ligamento cruzado anterior: correlação em ressonância magnética entre LCA, isquiotibiais, tendão patelar e tendão quadríceps[☆]

Fabiano Kupczik^a, Marlus Eduardo Gunia Schiavon^b, Bruno Sbrissia^b,
Rodrigo Caldonazzo Fávaro^b e Rafael Valério^{c,*}

R E S U M O

Objetivos: Mensurar em exames de ressonância magnética (RM) o tamanho da origem, a inserção e o comprimento do ligamento cruzado anterior (LCA) e seus possíveis enxertos para cirurgia de reconstrução em caso de lesão. Além desse, fez-se o cruzamento estatístico entre os dados para testar a hipótese de relação proporcional entre essas medidas anatômicas.

Materiais e métodos: Foram feitos 52 exames de RM entre 2008 e 2011 e avaliados de maneira aleatória em um estudo epidemiológico longitudinal retrospectivo. Para a mensuração da largura do LCA foi usado o corte coronal oblíquo, para o comprimento o corte sagital, para a inserção tibial o corte coronal e para a inserção femural o corte coronal oblíquo.

Que pergunta este gráfico lhe motiva?

Tópicos em
resultados
preliminares

Felipe
Figueiredo

Análise
Exploratória

EDA

Tabelas

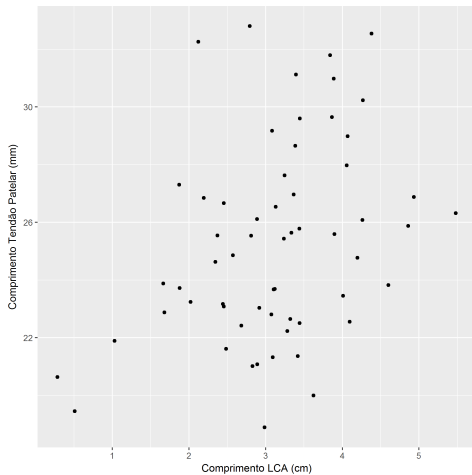
Figuras

Exercício

Resumo

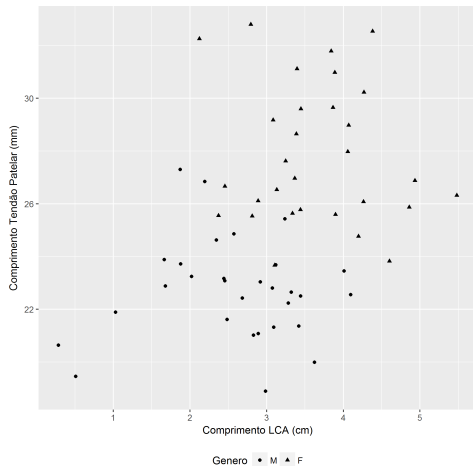
Referências

Aprofundamento



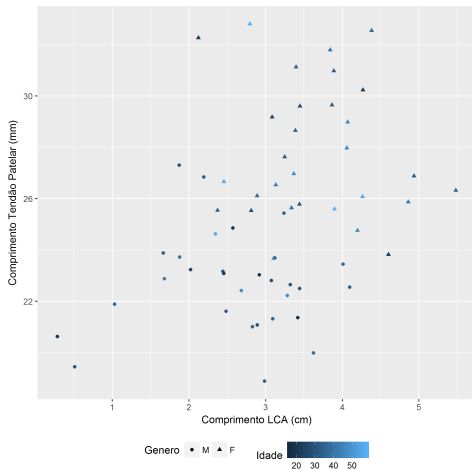
dados simulados com base no artigo

O Gênero parece influenciar?



dados simulados com base no artigo

E a idade?



dados simulados com base no artigo

1 Análise Exploratória

- EDA
- Tabelas
- Figuras
- Exercício
- **Resumo**
- Referências

2 Aprofundamento

- Aprofundamento

- 1 Não há escolha entre exploratória OU confirmatória – ambas são importantes
- 2 É preciso pensar em ciência no sentido amplo, e não no paradigma linear
- 3 Para uma confirmação adequada, precisamos de um desenho cuidadosamente randomizado
- 4 Pensar em exploratória como uma atitude, não apenas como um conjunto de técnicas – e usá-la antes da confirmatória

Tukey, 1980

1 Análise Exploratória

- EDA
- Tabelas
- Figuras
- Exercício
- Resumo
- **Referências**

2 Aprofundamento

- Aprofundamento

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

1 Análise Exploratória

- EDA
- Tabelas
- Figuras
- Exercício
- Resumo
- Referências

2 Aprofundamento

- Aprofundamento

Leitura obrigatória

Não há.

Leitura recomendada

Tukey (1980), We need both exploratory and confirmatory,
<http://www-ece.rice.edu/~fkl/classes/ELEC697/TukeyEDA.pdf>
(Acessado em 10/09/2015)