

# MAD-CB Aula 1: Introdução

James R. Hunter

7 de fevereiro de 2017

# MAD-CB

Figure 1:

**Matéria de Análise de Dados - Ciências Biomédicas**

# Os Quatro Pilares da Matéria

- Os conceitos principais de *estatística*

# Os Quatro Pilares da Matéria

- Os conceitos principais de *estatística*
- A organização, limpeza e análise prática dos dados

# Os Quatro Pilares da Matéria

- Os conceitos principais de *estatística*
- A organização, limpeza e análise prática dos dados
- As ferramentas de *programação e informática* que apoiarão o manuseio e análise dos dados

# Os Quatro Pilares da Matéria

- Os conceitos principais de *estatística*
- A organização, limpeza e análise prática dos dados
- As ferramentas de *programação e informática* que apoiarão o manuseio e análise dos dados
- Desenvolvimento de um *projeto particular* (ou pode ser em duplas) de pesquisa quantitativa

# Quem Sou Eu - Jim Hunter

- Pesquisador no Laboratório de Retrovirologia sob Prof. Dr. Ricardo Diaz
- Doutorando em Doenças Infecciosas
- Bacharel e Mestrado de Yale University
- Ensino
  - ▶ U. of Birmingham UK
  - ▶ U. of Michigan
  - ▶ École Nationale d'Administration Publique, Québec
  - ▶ Hunter College, City University of New York
  - ▶ Várias Escolas de MBA em São Paulo
- Mudou para Brasil em 1999

- Inglês é a idioma primária da ciência e da estatística
  - ▶ Publicações
  - ▶ Conferências
- Vou dar as palestras em português, **MAS**
  - ▶ Vou usar as palavras mais comuns para termos técnicos
- Qualquer duvida, **pergunte!**
- Textos são todos em inglês
- Estou usando esta matéria para desenvolver o conteúdo de um texto de data science biomédica em português



# Esta Matéria É Um Curso de Estatística?

- Sim e Não

- ▶ **Sim:** Nós vamos aprender estatística
- ▶ Aplicada às ciências biomédicas
- ▶ **Não:** Escutando palestras sobre tópicos básicos é um desperdício de tempo
- ▶ Foco mais em problemas de “data science” e pesquisa prática nas áreas biomédicas

# Então, Como????

- Aqui na sala de aula comigo

# Então, Como????

- Aqui na sala de aula comigo
- Vídeos adicionais no YouTube

# Então, Como????

- Aqui na sala de aula comigo
- Vídeos adicionais no YouTube
- Fazendo exercícios

# Então, Como????

- Aqui na sala de aula comigo
- Vídeos adicionais no YouTube
- Fazendo exercícios
- Usando as unidades de estatística de Khan Academy

# Então, Como????

- Aqui na sala de aula comigo
- Vídeos adicionais no YouTube
- Fazendo exercícios
- Usando as unidades de estatística de Khan Academy
- Outros cursos on-line

# Então, Como????

- Aqui na sala de aula comigo
- Vídeos adicionais no YouTube
- Fazendo exercícios
- Usando as unidades de estatística de Khan Academy
- Outros cursos on-line
- Livros recomendados

# Documentos da Matéria

- Todos os documentos, slides e programas da matéria ficam num repositório no GitHub
- Inclusive estes slides
- <https://github.com/jameshunterbr/MAD-CBt1>

The screenshot shows the GitHub repository page for `jameshunterbr / MAD-CBt1`. At the top, there are navigation tabs for `Code`, `Issues` (0), `Pull requests` (0), `Projects` (0), `Wiki`, `Pulse`, `Graphs`, and `Settings`. Below the navigation bar, the repository description reads "Arquivos, programas, etc. para Matéria de Análise de Dados - Ciências Biomédicas Turma 1". A `New` button and "Add topics" link are visible. The repository statistics bar shows 3 commits, 1 branch, 0 releases, 1 contributor, and the GPL-2.0 license. Below this, there are buttons for "Branch: master", "New pull request", "Create new file", "Upload files", "Find file", and "Clone or download". The file list shows the following files and their commit history:

File	Commit Message	Time
<code>.gitignore</code>	edit recursos	2 days ago
<code>LICENSE</code>	Initial commit	2 days ago
<code>MAD-CBt1.Rproj</code>	initial files	2 days ago
<code>README.md</code>	Initial commit	2 days ago



# Vídeos no YouTube

- <https://www.youtube.com/channel/UCbvgZ8RYeTtgjhAKE-jub5A>
- Canal em meu nome
- Olhe na lista de uploads. Os vídeos estarão lá

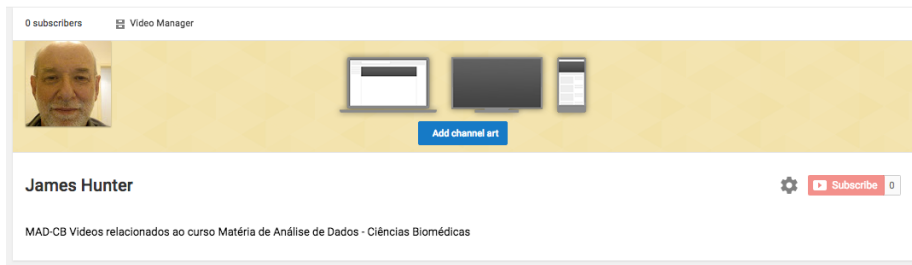


Figure 3:

- Impressionante site baseado nos EUA para ensino de assuntos técnicos
- Maioria das aulas já traduzidos em português
  - ▶ Graças a Fundação Lemann
- De Graça — FREE
- Nós vamos fazer as unidades de estatística de ensino médio
  - ▶ Mesmo que um curso primeiro de estatística

- <https://pt.khanacademy.org>

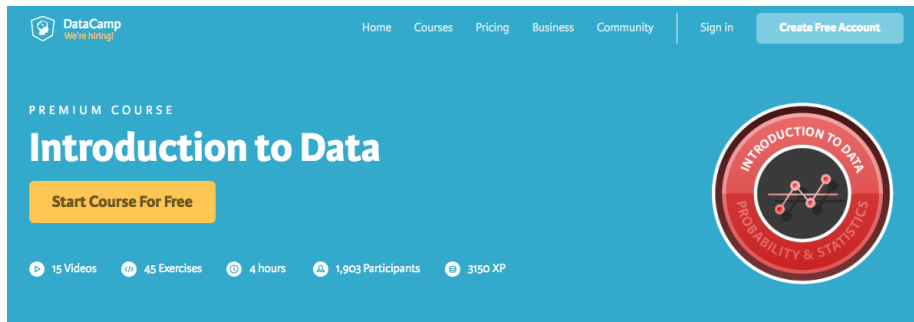


# Como Usar Khan Academy

- Criei um curso lá para monitorar seu progresso
- Para se inscrever no curso e fazer as unidades:
  - 1 Criar uma conta na Khan Academy
  - 2 Confirmar sua conta através de seu email
  - 3 Clique no seu nome no canto superior direita
  - 4 Clique na palavra Perfil
  - 5 Clique na aba Tutores
  - 6 Na caixa de Adicionar um professor, insira o código 4QY68Y
  - 7 Get to work!

# Alternativas a Khan Academy

- Datacamp curso em estatística básica
- Pago (oops)
- em inglês
- <https://www.datacamp.com/courses/introduction-to-data>



The screenshot shows the DataCamp website interface. At the top, there's a navigation bar with links for Home, Courses, Pricing, Business, and Community, along with Sign In and a Create Free Account button. The main section features the DataCamp logo and the text 'We're hiring!'. Below this, it says 'PREMIUM COURSE' and 'Introduction to Data'. A prominent yellow button says 'Start Course For Free'. At the bottom of the course card, there are icons and text indicating: 15 Videos, 45 Exercises, 4 hours, 1,903 Participants, and 3150 XP. On the right side of the course card, there is a circular graphic with the text 'INTRODUCTION TO DATA' and 'PROBABILITY & STATISTICS' around a central network diagram.

## Course Description

Scientists seek to answer questions using rigorous methods and careful observations. These observations—collected from the likes of field notes, surveys, and experiments—form the backbone of a statistical investigation and are called data. Statistics is the study of

Instructor(s):



James R. Hunter

MAD-CB Aula 1: Introdução

7 de fevereiro de 2017

13 / 30

- Como organizar e preparar os dados para a análise
- Uso de Excel como organizador
- Tipos de análises e modelos que podemos construir
  - ▶ Regressão
  - ▶ Análise de variância (ANOVA)
  - ▶ Machine Learning (supervised & unsupervised)
  - ▶ Análises especializadas para estudos biológicos
- Preparando estudos para a publicação
- Assunto primária das aulas

- Um derivativo open-source de linguagem “S”
  - ▶ Desenvolvido pelo antigo Bell Labs
- Primeiro versão: 1999
- Agora, Versão 3.3.2
- Comunidade ativa de desenvolvedores e programadores
- >10.000 pacotes/módulos publicados pela comunidade
- Ferramenta mais popular para modelos de estatística na academia hoje
- Sem custo/GRATIS/FREE

# R É Uma Linguagem Não Um Software Completo

- Modelo diferente dos softwares tradicionais
  - ▶ SPSS
  - ▶ Graphpad Prism
  - ▶ SAS
  - ▶ Statistica
  - ▶ STATA



- Sistemas de menus e caixas escondem o código que os sistemas escrevem

# Problemas com os Tradicionais

- Sistemas de menus e caixas escondem o código que os sistemas escrevem
- Difícil saber se você pode reproduzir exatamente a seqüência de comandos

# Problemas com os Tradicionais

- Sistemas de menus e caixas escondem o código que os sistemas escrevem
- Difícil saber se você pode reproduzir exatamente a seqüência de comandos
- Custo frequentemente absurdo

- Sistemas de menus e caixas escondem o código que os sistemas escrevem
- Difícil saber se você pode reproduzir exatamente a seqüência de comandos
- Custo frequentemente absurdo
- “*You can only do what the buttons say you can do.*” — Sacha Epskamp, U de Amsterdã<sup>1</sup>

---

<sup>1</sup>Baker, Monya, “Code Alert”, **Nature**, Vol 541, 26/1/2017, p. 563 - 565.

- Sistemas de menus e caixas escondem o código que os sistemas escrevem
- Difícil saber se você pode reproduzir exatamente a seqüência de comandos
- Custo frequentemente absurdo
- *“You can only do what the buttons say you can do.”* — Sacha Epskamp, U de Amsterdã<sup>1</sup>
- *“Você pode fazer somente o que os programadores dos botões acham que o botão faz.”* — Jim Hunter, UNIFESP

---

<sup>1</sup>Baker, Monya, “Code Alert”, **Nature**, Vol 541, 26/1/2017, p. 563 - 565.

# Com R, Você Tem Controle de Resultados

- Com um pouco treinamento, pode depender que os resultados seriam confiáveis
- Você escreve uma serie de instruções (código) para dirigir as operações do programa
- Você controle, outros podem facilmente auditar
  - ▶ Chave para “reproducible research”
- Curva de aprendizagem
  - ▶ Inicial – suave
  - ▶ Avançado – mais inclinada

# Quase Toda a Matéria do Curso Está Sendo Preparada com R

- Estes slides
- Handouts
- Programas
- Gráficos
- Com ajuda de IDE para R, RStudio

# Projeto Final da Matéria

- Divisão da turma em até 10 grupos
  - ▶ De 1 até 3 pessoas
- Escolha de uma pesquisa para analisar
  - ▶ Pode ser um projeto de tese ou do laboratório
  - ▶ Nós podemos ajudar você com a escolha
- Grupo vai fazer uma apresentação sobre a análise quantitativa dos dados relacionados a tema do projeto
  - ▶ Começando ao final do março
  - ▶ Ordem seria sorteada na aula
- Na aula final, grupo vai submeter um relatório resumindo a análise que fez
  - ▶ Técnicas e ferramentas usadas
  - ▶ Resultados
  - ▶ Quais técnicas funcionaram bem, quais não tão bem
  - ▶ No estilo de um publicação formal no jornal científico



- Altamente recomendados!
- Ajuda com problemas de programação
- Projetos
- Fazendo problemas

- Vai ter
- **MAS** são diagnósticos, não avaliações formais
- Provas como experiência de aprendizagem

- Têm em Khan Academy
- Vou criar conjuntos de problemas usando um pacote de R – `swirl`
- Faz parte de nota final

- Jim Hunter estará disponível quintas de 13.30h até 16.00h
  - ▶ Ed. de Pesquisa 2, Pedro de Toledo, 669, 6º Andar fundos
- Para reforçar o que estão aprendendo na aula, na Khan Academy, etc.
- Tirar dúvidas sobre o conteúdo e a administração da matéria
- Não precisa marcar antecipadamente; **Just show up**
- Email do Jim: [jhunter@unifesp.br](mailto:jhunter@unifesp.br)
- Cel do Jim (para Whatsapp): 11-9-5327-5656

- Crawley, **Statistics: An Introduction Using R**, (Wiley)

- Crawley, **Statistics: An Introduction Using R**, (Wiley)
- Diez, Barr & Cetinkaya-Rundel, **OpenIntro Statistics**, (<http://openintro.org>)

- Crawley, **Statistics: An Introduction Using R**, (Wiley)
- Diez, Barr & Cetinkaya-Rundel, **OpenIntro Statistics**, (<http://openintro.org>)
- Navarro, **Learning statistics with R: A tutorial for psychology students and other beginners**, (<http://learningstatisticswithr.com>)

- Crawley, **Statistics: An Introduction Using R**, (Wiley)
- Diez, Barr & Cetinkaya-Rundel, **OpenIntro Statistics**, (<http://openintro.org>)
- Navarro, **Learning statistics with R: A tutorial for psychology students and other beginners**, (<http://learningstatisticswithr.com>)
- Nolan & Speed, **Stat Labs: Mathematical Statistics through Applications**, Springer



- Crawley, **Statistics: An Introduction Using R**, (Wiley)
- Diez, Barr & Cetinkaya-Rundel, **OpenIntro Statistics**, (<http://openintro.org>)
- Navarro, **Learning statistics with R: A tutorial for psychology students and other beginners**, (<http://learningstatisticswithr.com>)
- Nolan & Speed, **Stat Labs: Mathematical Statistics through Applications**, Springer
- Vickers, **What is a P-Value Anyway?**, Addison-Wesley

- Kabacoff, **R in Action: Data analysis and graphics with R, 2e** (Manning)

- Kabacoff, **R in Action: Data analysis and graphics with R**, 2e (Manning)
- Peng, **R Programming for Data Science** (Leanpub)

- Kabacoff, **R in Action: Data analysis and graphics with R**, 2e (Manning)
- Peng, **R Programming for Data Science** (Leanpub)
- Peng, Kross & Anderson, **Mastering Software Development in R** (Leanpub)

- Kabacoff, **R in Action: Data analysis and graphics with R**, 2e (Manning)
- Peng, **R Programming for Data Science** (Leanpub)
- Peng, Kross & Anderson, **Mastering Software Development in R** (Leanpub)
- Phillips, **YaRrr!: The Pirate's Guide to R** (<http://www.thepiratesguidetor.com>)

- Kabacoff, **R in Action: Data analysis and graphics with R**, 2e (Manning)
- Peng, **R Programming for Data Science** (Leanpub)
- Peng, Kross & Anderson, **Mastering Software Development in R** (Leanpub)
- Phillips, **YaRrr!: The Pirate's Guide to R** (<http://www.thepiratesguidetor.com>)
- Wickham & Grolemund, **R for Data Science**, (<http://r4ds.had.co.nz> or O'Reilly)

- Kabacoff, **R in Action: Data analysis and graphics with R**, 2e (Manning)
- Peng, **R Programming for Data Science** (Leanpub)
- Peng, Kross & Anderson, **Mastering Software Development in R** (Leanpub)
- Phillips, **YaRrr!: The Pirate's Guide to R** (<http://www.thepiratesguidetor.com>)
- Wickham & Grolemund, **R for Data Science**, (<http://r4ds.had.co.nz> or O'Reilly)
- Zumel & Mount, **Practical Data Science with R** (Manning)

- Hansen, **Bioconductor: An Introduction to Core Technologies** (Leanpub)
- Irizarry & Love, **Data Analysis for the Life Sciences** (Leanpub)

**Livros de Leanpub:** <https://leanpub.com>



## 5 Livros que Deve Ler Porque São Bons

- Leonard Mlodinow, **O Andar do Bêbado**

## 5 Livros que Deve Ler Porque São Bons

- Leonard Mlodinow, **O Andar do Bêbado**
- David Salsburg, **Uma Senhora Toma Chá**

## 5 Livros que Deve Ler Porque São Bons

- Leonard Mlodinow, **O Andar do Bêbado**
- David Salsburg, **Uma Senhora Toma Chá**
- Ian Stewart, **17 Equações que Mudaram o Mundo**

## 5 Livros que Deve Ler Porque São Bons

- Leonard Mlodinow, **O Andar do Bêbado**
- David Salsburg, **Uma Senhora Toma Chá**
- Ian Stewart, **17 Equações que Mudaram o Mundo**
- Peter L. Bernstein, **Desafiando os Deuses: A História do Risco**

# 5 Livros que Deve Ler Porque São Bons

- Leonard Mlodinow, **O Andar do Bêbado**
- David Salsburg, **Uma Senhora Toma Chá**
- Ian Stewart, **17 Equações que Mudaram o Mundo**
- Peter L. Bernstein, **Desafiando os Deuses: A História do Risco**
- Randall Munroe, **E Se?: Respostas Científicas para Perguntas Absurdas**

- Precisa contar para mim o que vocês querem aprender

- Precisa contar para mim o que vocês querem aprender
- Precisa me avisar se estou indo rápido demais ou devagar demais

- Precisa contar para mim o que vocês querem aprender
- Precisa me avisar se estou indo rápido demais ou devagar demais
- Vou tentar deixar cálculo e álgebra linear fora (mas, não é sempre possível)



- Precisa contar para mim o que vocês querem aprender
- Precisa me avisar se estou indo rápido demais ou devagar demais
- Vou tentar deixar cálculo e álgebra linear fora (mas, não é sempre possível)
- Vou usar dados de nossa área quando for possível – se tiver alguns datasets interessantes, me avise

- Pergunte bastante; participação ajuda todos
- Se você fica com uma pergunta, pode garantir que outra pessoa tem a mesma pergunta
- Não existem perguntas burras

*There are naive questions, tedious questions, ill-phrased questions, questions put after inadequate self-criticism. But every question is a cry to understand the world. **There is no such thing as a dumb question.***

- Carl Sagan, **The Demon-Haunted World: Science as a Candle in the Dark**, p. 303.