Bioinformática Aplicada – 2024

Modulo R

James R. Hunter, D.Sc. Laboratório de Retrovirologia, UNIFESP

2024-10-09

1 Introdução

Nas seis semanas do modulo, nós vamos aprender a linguagem de computação R. Esta é uma linguagem muito utilizada hoje em dia para análises estatísticas em biologia e medicina. Nos EUA, é a linguagem e ferramenta mais utilizada nas faculdades de ciências.

Ao final das aulas, vocês não tornarão especialistas na linguagem, mas terão uma base firme para aprender técnicas avançadas sozinho ou em cursos avançados. Durante as aulas, vou compartilhar com vocês várias matérias para orientar suas pesquisas futuras. Entre elas serão uma bibliografia dos livros, web sites, blogs que vale a pena consultar.

Nas suas carreiras, vocês encontrarão um monte dos dados nos artigos, nas teses e mesmo nos artigos dos jornais populares. Temos acesso hoje aos rios dos dados. Como profissionais nas áreas de saúde e medicina, precisamos saber como analisar esses dados e separar os estudos que comuniquem informação médica e biológica convincente dos outros que são lixo ou, pior, “fake”. Precisamos também construir nossos próprios conjuntos dos dados e conduzir nossas próprias análises. Aqui, R é uma ferramenta excelente.

Este módulo ajudará vocês com as habilidades necessárias de computação e de programação que são essenciais para você possa organizar seus dados, os analisar e apresentar suas conclusões. Não é um curso tradicional teórico de estatística. Vai mostrar para vocês como aplicar as técnicas de estatística que já conhecem (ou vão aprender) via programação invés de utilização dos grandes programas de estatística. Vai mostrar para você como utilizar efetivamente ferramentas de programação que permitem que vocês constroem resultados científicos robustos.

Principalmente, aprenderemos por fazer. Por causa do pequeno tamanho do grupo, vamos fazer muitos exercícios “*live-coding*” na sala. Até na primeira aula, você vai fazer uma análise útil. Cada semana, suas habilidades crescerão em escopo (com mais técnicas) e aprofundarão. Começamos com análises bioestatísticas básicas que ajuda com a descrição dos dados e progredimos para utilizar técnicas de bioestatística e aprendizado de máquina supervisionado[1] para alcançar conclusões dos dados de amostras sobre as populações que as amostras representam.

---

[1] Uma palavra que quer dizer o modelo terá uma variável dependente. Não se preocupe com esta linguagem técnica agora. Na semana quatro, você entenderá bem o conceito.

O curso não coloca o foco na matemática teórica. Você não precisa uma formação forte em matemática. Você vai precisar saber algumas ferramentas básicas de matemática como o conceito de uma soma (∑), logaritmos e exponentes, e a equação de uma linha reta. Na primeira aula, faremos uma revisão de alguns desses conceitos.

2 Porque Programação – Porque R

Programação das análises dos dados com uma linguagem permite que podemos construir nossas análises na forma de uma receita que você, seu computador, e qualquer leitor pode seguir e reproduzir. As análises terão exatamente os parâmetros que necessitam para fazer a análise certa. As grandes softwares integrados de estatística como SPSS ou GraphPad Prism não permitem isso e frequentemente um analista pode mudar os resultados inesperadamente por clicar uma caixa ou mudar um parâmetro da análise sem perceber.

Outra vantagem é que R e RStudio são softwares "*open-source*" que são disponíveis **de graça** no internet.

3 Biografia - James R. Hunter, D.Sc.

Dr. Hunter fez o doutorado em Doenças Infecciosas em UNIFESP em 2019. Antes, ele tinha feito um B.A. e um M.C.P. (Mestrado em Urbanismo), ambos da Universidade Yale nos EUA. Desde 1970, ele ensina matérias sobre métodos quantitativos, estatística e pesquisa operacional nos EUA, Inglaterra, Canadá e o Brasil. Atualmente, ele é Professor Afiliado da UNIFESP e um Pós-Doutorando no Laboratório de Retrovirologia da Escola Paulista de Medicina. Ele mora em Brasil desde 1999 e iniciou atividades em UNIFESP em 2014.

4 Matérais para Aprender R

Mais importante, nenhum curso vai ensinar tudo sobre um tópico. É assim com este curso. Precisa consultar outras matériais que tratam do assunto também. Esses matérias podem apresentar o assunto que você está estudando com uma visão diferente e mais convincente para você.

Todos os materiais usados nas aulas serão depositados no repositório do GitHub: bioapp2024_r (https://github.com/jameshunterbr/bioapp2024_r). Para os alunos inscritos formalmente no curso também existe uma área no Google Classroom.

Abaixo é uma lista dos livros que alunos achariam úteis para o curso. A maioria é **de graça**.

4.1 Bibliografia — Estatística

- Diez, Barr & Cetinkaya-Rundel, **OpenIntro Statistics 4**, (http://openintro.org)

- Navarro, D. **Learning statistics with R: A tutorial for psychology students and other beginners**, (http://learningstatisticswithr.com)

4.2 Bibliografia — R, Programação e Análise dos Dados

- Ismay & Kim, **Statistical Inference via Data Science: A ModernDive into R and the Tidyverse** (https://moderndive.com)
- Irizarry, **Introduction to Data Science** (https://rafalab.github.io/dsbook)
- Irizarry & Love, **Data Analysis for the Life Sciences** (Leanpub)
- Peng, **R Programming for Data Science** (Leanpub & Bookdown)
- Wickham & Grolemund, **R for Data Science**, (http://r4ds.had.co.nz ou O'Reilly)

Livros de Leanpub: https://leanpub.com; Livros de Bookdown: https://bookdown.org/

4.3 5 Livros que Deve Ler sobre Análise dos Dados (Porque São Bons Demais)

Agora algumas livros que você pode curtir bastante e que farão você muito mais confortável com a matemámatica.

- 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** (e o novo volume 2)

Além desses livros, vou indicar uma serie dos site, blogs, YouTubes, antes e depois das aulas para ajudar vocês entenderem melhor como aproveitar de R.

5 Programa das Aulas

Aulas e Salas

Data	Aula
27 Outubro	Conceitos Básicos de Programação e de R; Tipos de Dados
05 Novembro	Fluxo de Controle; Condicionais; Funções; Bibliotecas
12 Novembro	Visualização dos Dados com Base R e ggplot
19 Novembro	Testes de Dados em R com 2 ou Mais Variáveis
26 Novembro	Bioinformática em R; Bioconductor

Data	Aula
02 Dezembro	Funções e Pacotes

6 Laptops e O Que Precisa Fazer para Preparar para a Primeira Aula

Tragam seus laptops para as aulas. Vamos usar eles para programar exercícios ao vivo. Se não puder, precisa arrumar um parceiro entre os alunos com quem possa compartilhar o uso do aparelho.

Os laptops devem ter instalados o programa de **R** e o IDE ("Integrated Development Environment") para R, chamado **RStudio**. O site onde achar R é CRAN: https://cran.r-project.org/. O site de RStudio é https://posit.co/products/open-source/rstudio/.

6.1 Pensamento final

Todos os documentos deste modulo serão feitos em R, utilizando normalmente o sub-sistema Quarto. Inclusive esta documento. Slides, documentos, gráficos, todos. Acho que vão achar R um sistema poderoso para gerenciamento dos seus projetos analiticos.