Bio 204: Biological Data Analysis

Paul M. Magwene

Department of Biology

29 August 2016

Welcome

- Introductions
- What is “Biological Data Analysis”?
- Grading and course policies
- Course Overview

Teaching Team

Instructor

- Paul Magwene – Associate Professor, Department of Biology; Director of Graduate Program in Computational Biology and Bioinformatics

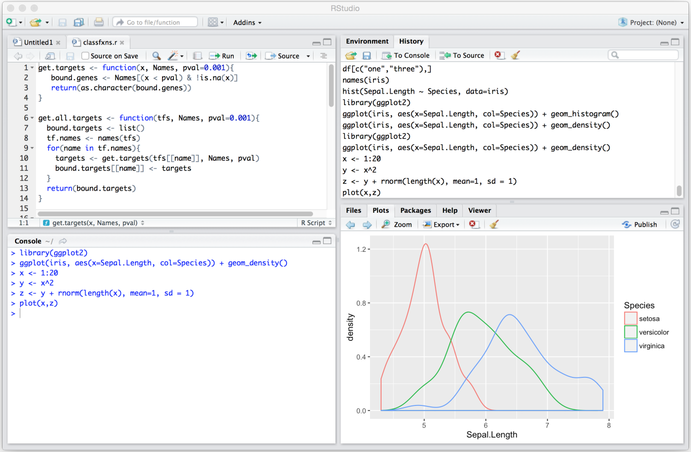
TA

- Cullen Roth – Graduate student in the University Program in Genetics and Genomics. Extensive mathematical and statistical computing experience.

What is “Biological Data Analysis”?

- Scientific Computing
    - Data visualization, exploration, description
    - Data “munging” – converting, combining, filtering, subsetting, and restructuring complex data
    - Reproducible computational research
    - Simulation
- Statistics – the science of learning from data
    - Classic parametric and non-parametric methods – t-tests, ANOVA, regression, etc
    - Machine learning – clustering, classification, dimensionality reduction, etc

Computing Environment: R / RStudio



```
get.targets <- function(x, Names, pval=0.001){
   bound.genes <- Names[(x < pval) & !is.na(x)]
   return(as.character(bound.genes))
}

get.all.targets <- function(tfs, Names, pval=0.001){
  bound.targets <- list()
  tf.names <- names(tfs)
  for(name in tf.names){
    targets <- get.targets(tfs[[name]], Names, pval)
    bound.targets[[name]] <- targets
  }
  return(bound.targets)
}
```

```
df[c("one","three"),]
names(iris)
hist(Sepal.Length ~ Species, data=iris)
library(ggplot2)
ggplot(iris, aes(x=Sepal.Length, col=Species)) + geom_histogram()
ggplot(iris, aes(x=Sepal.Length, col=Species)) + geom_density()
library(ggplot2)
ggplot(iris, aes(x=Sepal.Length, col=Species)) + geom_density()
x <- 1:20
y <- x^2
z <- y + rnorm(length(x), mean=1, sd = 1)
plot(x,z)
```

```
> library(ggplot2)
> ggplot(iris, aes(x=Sepal.Length, col=Species)) + geom_density()
> x <- 1:20
> y <- x^2
> z <- y + rnorm(length(x), mean=1, sd = 1)
> plot(x,z)
> 
```

Figure: The RStudio Environment

Syllabus, First Third

- Getting up to speed with R
- Data visualization and exploration
- Quantitative measures for describing univariate and bivariate data
- Regression and curve fitting

Syllabus, Second Third

- Probability
- Statistical distributions
- Understanding sampling distributions of statistics of interest through statistical simulation
- Central Limit Theorem

Syllabus, Last Third

- Confidence Intervals
- Hypothesis testing and statistical power
- t-tests
- ANOVA
- Regression revisited
- χ^2 and contingency tables

Course policies: Academic Integrity

- All students are expected to adhere to and have an obligation to act in accordance with the Duke Community Standard.
- Strict adherence to the plagiarism policy described in the Community Standard will be observed. Any violations of the community standard will be referred to the undergraduate judicial board.
- Students are encouraged to study together and discuss the course material.

Course policies: Missed classes

- Religious/Athletic/Interviews – Must notify instructor at least one week in advance about missed class time.
- Illness – STINF or letter from academic dean if long-term illness.
- Students with excused absences other than illness are still expected to submit problem sets by assigned dates.

Grading

Quizzes

In-class quizzes related to readings and lecture material from previous classes. Multiple choice or short answer.

Problem sets

Weekly statistical and computational problems based on the material covered in lectures and the readings. 12 assignments total; lowest score dropped.

Late assignments

Homework assignments that are submitted late without a STINF or instructor approval will receive half credit if submitted within 24 hours of the due date, or zero credit thereafter.

Bonus points for on-time assignment completion

Students completing all problem sets and quizzes on time, and without any excused absences or STINFs, will receive bonus points towards their final grade.

Texts

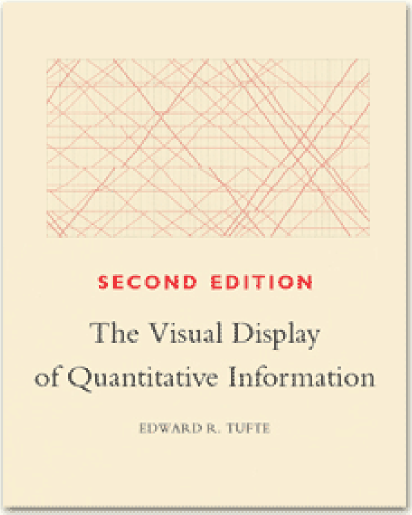
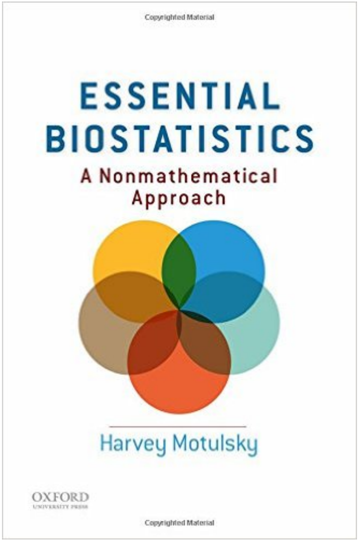


Figure: Tufte, 2001. The Visual Display of Quantitative Information.

Texts



ESSENTIAL BIOSTATISTICS

A Nonmathematical Approach

Harvey Motulsky

OXFORD
UNIVERSITY PRESS

Figure: Motulsky, 2015. Essential Biostatistics: A Nonmathematical Approach.

Texts

- Nature Methods, Points of Significance – A series of short articles, published 2013-2015, on key statistical topics, aimed at the working biologist.

Class materials

- Sakai – submitting problem sets and viewing grades
- Class wiki – everything else. See link in the PDF version of this slide or on Sakai.
    - Direct link:

  https://github.com/Bio204-class/Bio204-Fall-2016/wiki

In class survey

Fill out the survey at `https://goo.gl/forms/iQiH1ml08JNkMzgA2`

Hands-on exercise: Describe a small data set

See handout