Introduction to An Introduction to Computational Data Analysis for Biology

http://jarrettbyrnes.info/biol697

Jarrett Byrnes

UMass Boston

Sept 4, 2012

Outline for Today

1. Why this course?
2. Who are we?
3. How will we approach the work?
4. How will this course work?
5. R!

What is this Course About?

- Introduction to - starting with the basics
- Computational - programming & other computational tools
- Data - collection, curation, maintenance of information
- Analysis - statistics
- for Biology - SCIENCE FIRST

What I want for you:l

To be able to go from your ideas about a system to a model, fit and evaluated with the appropriate data.

Course Goals

1. Learn how to think about your research in a systematic way to design efficient observational & experimental studies.
2. Understand how to get the most bang for your buck from your data.
3. Make you effective collaborators with statisticians.
4. Make you comfortable enough to learn and grow beyond this class.

Why a Computational Focus?

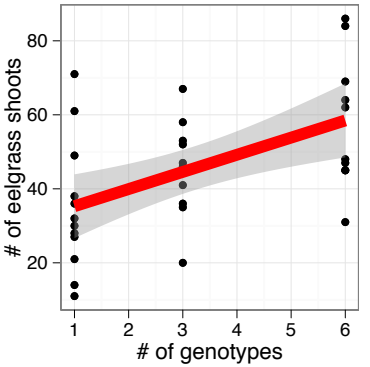
```
library(plyr)

d_ply(eelgrass, .genotypes, function(x) {
    print(summary(lm(shoots ~ geese, data = x)))
})
```

- Programming is a necessary skill for everything
- We live in the era of big data
- Comfort with algorithmic thinking helps your science

Introduction to An Introduction to Computational Data Analysis for Biology

How will we use statistics?

- Estimation
  - Parameter in model
  - Variance in parameter estimation
- Model Evaluation
  - What parameters should be included in a model?
  - Does a model fit the data?
  - Comparison of competing hypotheses



Introduction to An Introduction to Computational Data Analysis for Biology

Two Different Skillsets

- Statistics
- Programming

Introduction to An Introduction to Computational Data Analysis for Biology

Questions?

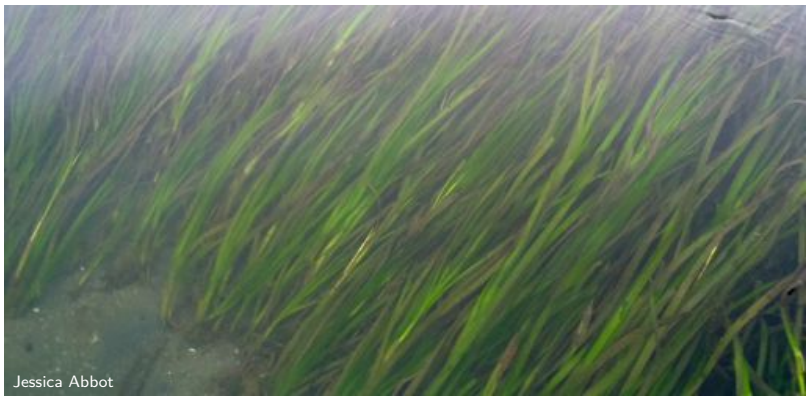
Who Are You?

1. Name
2. Lab
3. Brief research description
4. Why are you here?

Our Approach to Data Analysis

Data from Reusch et al. 2005 PNAS

Start with a Question

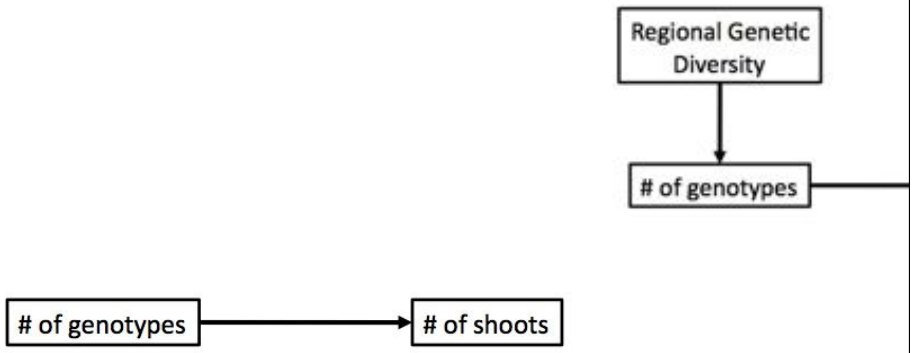


Does seagrass genetic diversity increase productivity?

Build an Understanding of the System

1. Literature
2. Observation
3. Natural History

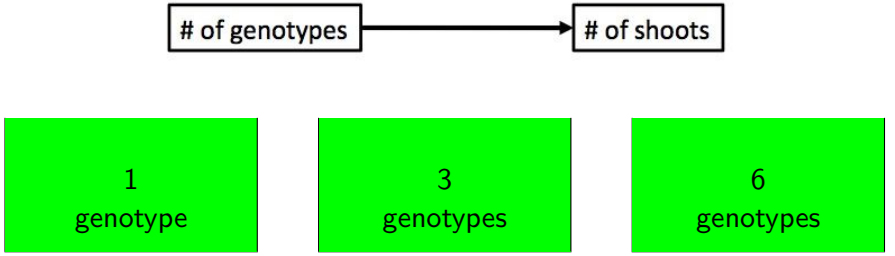
Construct a Model of the System



Causal Graph

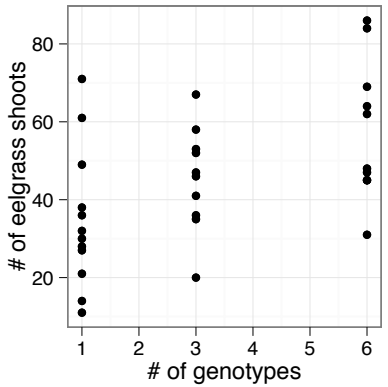
Introduction to An Introduction to Computational Data Analysis for Biology

Collect the Data to Best Estimate & Test the Model



Introduction to An Introduction to Computational Data Analysis for Biology

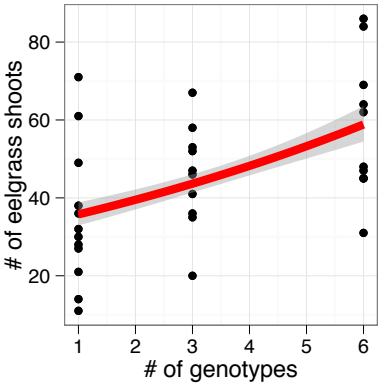
Look at Your Data



Fit a model(s), chosen to suit this data

Introduction to An Introduction to Computational Data Analysis for Biology

Analysis!



Build Open Reproducible Research

Many Methods of Sharing Data, Methods, and Results Beyond Publication

1. GitHub - public code repository
2. FigShare - share key figures, get a doi
3. Blog - open 'notebook'
4. Dryad or Other Repository - post-publication data sharing

Questions?

Lecture/Lab/Labinar?

- I will yammer on
- R lab will be part of class
- Notes available at `http://jarrettbyrnes.info/biol697`
- Slide source available at `http://github.com/jebyrnes/biol697`

Special Topics

Additional special topics mini-labinars, e.g. knitr & LaTeX

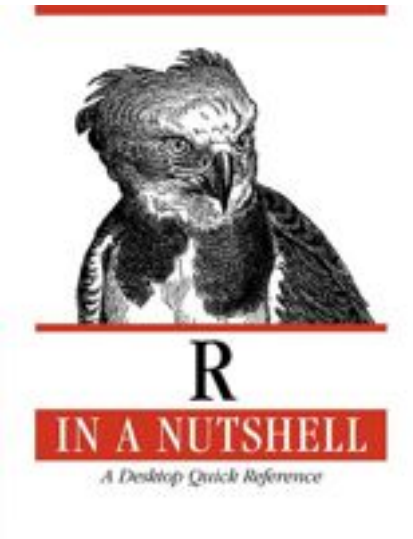
Readings for Class

Whitlock, W.C. and Schluter, D. (2008) The Analysis of Biological Data. Roberts and Company Publishers.

`http://www.zoology.ubc.ca/~whitlock/ABD/teaching/index.html`

Introduction to An Introduction to Computational Data Analysis for Biology

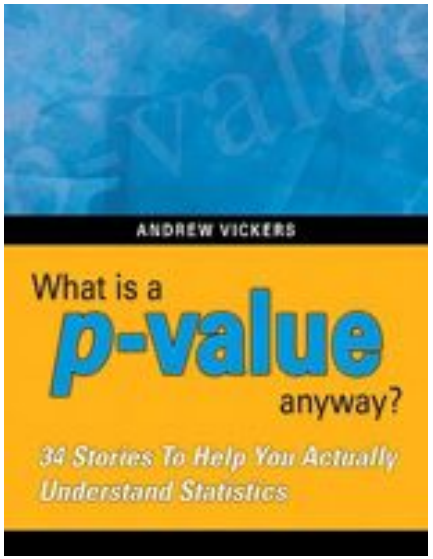
Readings for Class



Adler, J. (2009) R in a Nutshell: A Desktop Quick Reference. O'Reilly.

Introduction to An Introduction to Computational Data Analysis for Biology

Reflections



Media. Vickers, A. (2009) What is a p-value anyway? 34 Stories to Help You Actually Understand Statistics. Addison Wesley.

Write a weekly reflection. 1 page. Graded for participation (10%). 1 entry posted per week for discussion. `http://learningdata.wordpress.com/`

Introduction to An Introduction to Computational Data Analysis for Biology

Problem Sets

- 40% of your grade
- Adapted from Whitlock and Schluter
- Will often require R
- Turn in all code, and it must be understandable

Practical Exams

- 20% Midterm, 30% final
- Real world data analysis problems
- Will require R
- Turn in all code, and it must be understandable

Extra Credit: Your Work

- 10% Extra
- Report on your own data
- Cogently present what you did, why you did it, and the results & interpretation
- Data & Code must be accessible & understandable
- Extra points for putting work online so others can use & view your work

Topics

1. Data & Data Management
2. Biological Processes & Statistical Distributions
3. Data Visualization
4. Simulation & Basic Hypothesis Testing
5. Sampling Design
6. Fitting Linear Models: Least Squares
7. Fitting Linear Models: Likelihood
8. Generalized Linear Models
9. Experiments & the Linear Model (ANOVA)
10. Multiple Continuous Predictors
11. What should I sample? Simpson's Paradox
12. Interactions & Nonlinearities
13. Bootstrapping
14. Model Comparison

Questions?

What is R?

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- *an effective data handling and storage facility,*
- *a suite of operators for calculations on arrays, in particular matrices,*
- *a large, coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, and*
- *a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.*

From http://r-project.org

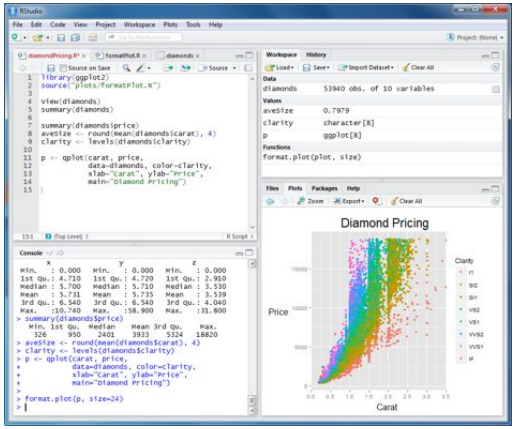
What is R?

- A programming language uniquely developed for statistical analysis

Why R?

1. Free
2. Huge growing community
3. Packages to do almost anything
4. Makes reusable research easy
5. C-based language
6. Syntax naturally matches analytical thinking

What is R Studio?



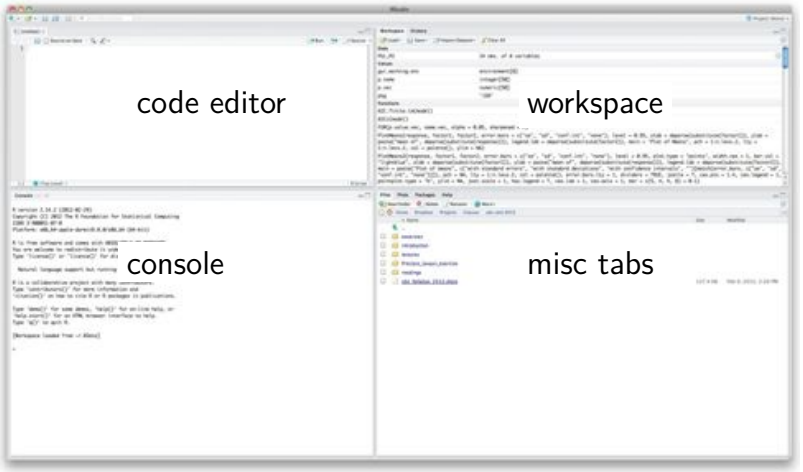
- Cross-Platform Graphical User Interface for R
- It is not R

Let's Fire It Up!

Open R-Studio.

Don't have it? Download it from `http://rstudio.org`

What do you see?



The Console and Math

```
1 + 1

## [1] 2
```


Everything is an Object

```
a.number <- 1 + 1
```

```
a.number

## [1] 2
```

Note: Comment Your Code as You Write with

The text after # is not evaluated.

```
# This is going to be the number two
a.number <- 1 + 1
```

```
##### ----------
# You can get creative with comments to separate code
# blocks And write a lot, which is good practice
##### ----------
```


Functions Work on Objects

```
sin(a.number)

## [1] 0.9093
```

How to get help for a function

```
`?`(cos)

help(cos)

`?`(`?`("cosine function"))
```


Lots of Object Types - like Data!

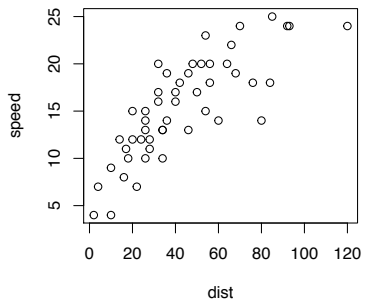
```
head(cars, n = 3)  #note the n= argument!

##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

Try looking at all of cars
Can be lots of information stored in an object

```
names(cars)

## [1] "speed" "dist"
```

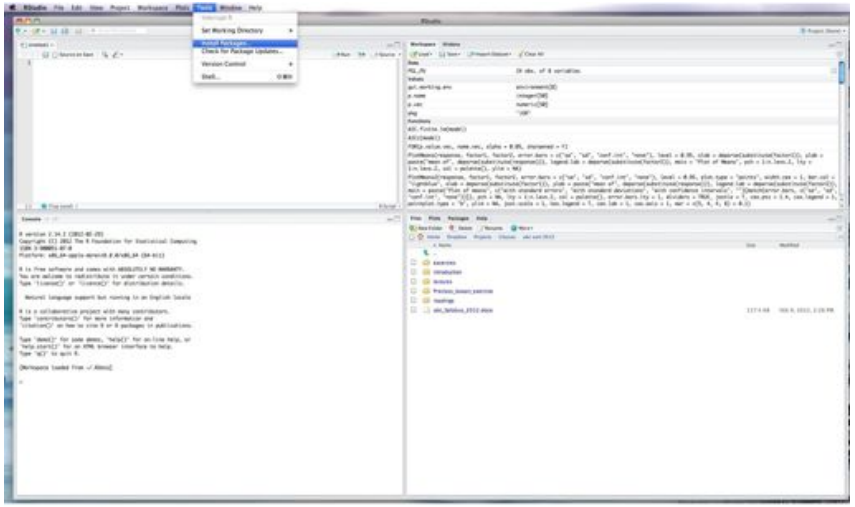

Graphics are a Snap

```
plot(speed ~ dist, data = cars)
```

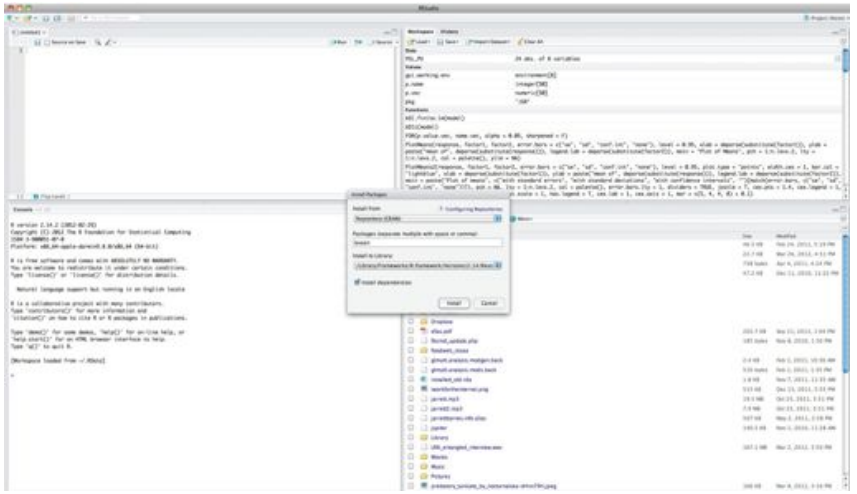


Look at ?plot to see other arguments to change appearance

Installing Packages



Installing Packages



Installing Packages

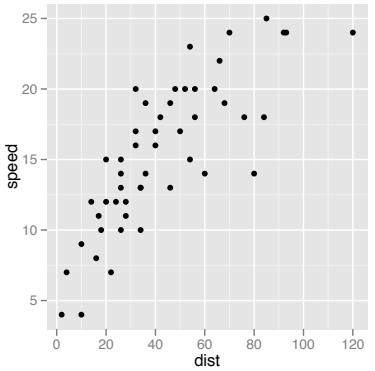
You can also install packages from the command line.

```
install.packages("ggplot2", repos = "http://cran.case.edu/",
    dependencies = TRUE)
```

Using one of the above methods, install the package ggplot2 and its dependencies now.

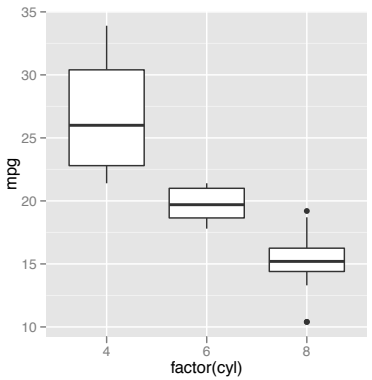
Using a Package

```
library(ggplot2)

qplot(dist, speed, data = cars)
```



You Try It

- Load ggplot2 and look at the mtcars data set
- Look at the qplot help file & demos
- Make two plots



Introduction to An Introduction to Computational Data Analysis for Biology

Questions?

Introduction to An Introduction to Computational Data Analysis for Biology

Next time

- Data Management!
- Contact me if you are not enrolled
- Read chapter 1 of the Nutshell
- Read P-Values chapters 1, 32-34 & ponder