Outline for Today

Introduction to An Introduction to Computational Data Analysis for Biology http://jarrettbyrnes.info/biol697 Jarrett Byrnes UMass Boston Sept 4, 2012	<ol> <li>Why this course?</li> <li>Who are we?</li> <li>How will we approach the work?</li> <li>How will this course work?</li> <li>R!</li> </ol>
Introduction to An Introduction to Computational Data Analysis for Biology	Introduction to An Introduction to Computational Data Analysis for Biology
What is this Course About?	Course Goals
<ul> <li>Introduction to - starting with the basics</li> <li>Computational - programming &amp; other computational tools</li> <li>Data - collection, curation, maintenance of information</li> <li>Analysis - statistics</li> <li>for Biology - SCIENCE FIRST</li> <li>What I want for you:I</li> <li>To be able to go from your ideas about a system to a model fit</li> </ul>	<ol> <li>Learn how to think about your research in a systematic way to design efficient observational &amp; experimental studies.</li> <li>Understand how to get the most bang for your buck from your data.</li> <li>Make you effective collaborators with statisticians.</li> <li>Make you comfortable enough to learn and grow beyond this</li> </ol>

#### Why a Computational Focus?

```
library(plyr)
d_ply(eelgrass, .genotypes, function(x) {
    print(summary(lm(shoots <sup>-</sup> geese, data = x)))
})
```

- Programming is a necessary skill for everything
- We live in the era of big data

Introduction to An Introduction to Computational Data Analysis for Biology

Comfort with algorithmic thinking helps your science

#### 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

# Questions?

	Who	Are	You?	
--	-----	-----	------	--

1. Name

2. Lab

3. Brief research description

Introduction to An Introduction to Computational Data Analysis for Biology

4. Why are you here?

# Our Approach to Data Analysis

Data from Reusch et al. 2005 PNAS

Introduction to An Introduction to Computational Data Analysis for Biology

#### Start with a Question



Does seagrass genetic diversity increase productivity?

#### Build an Understanding of the System

- 1. Literature
- 2. Observation
- 3. Natural History



Introduction to An Introduction to Computational Data Analysis for Biology	Introduction to An Introduction to Computational Data Analysis for Biology
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?
Introduction to An Introduction to Computational Data Analysis for Biology	Introduction to An Introduction to Computational Data Analysis for Biology
Lecture/Lab/Labinar?	Special Topics

#### Readings for Class

# XXX

The Analysis of Biological Data WHITLOCK SCHLUTER

Introduction to An Introduction to Computational Data Analysis for Biology

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

#### 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





34 Stories To Help You Actuall Understand Statistics 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/

#### 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

Introduction to An Introduction to Computational Data Analysis for Biology	Introduction to An Introduction to Computational Data Analysis for Biology
Topics	
<ol> <li>Data &amp; Data Management</li> <li>Biological Processes &amp; Statistical Distributions</li> <li>Data Visualization</li> <li>Simulation &amp; Basic Hypothesis Testing</li> <li>Sampling Design</li> <li>Fitting Linear Models: Last Squares</li> <li>Fitting Linear Models: Likelihood</li> <li>Generalized Linear Models</li> <li>Experiments &amp; the Linear Model (ANOVA)</li> <li>Multiple Continuous Predictors</li> <li>What should I sample? Simpson's Paradox</li> <li>Interactions &amp; Nonlinearities</li> <li>Bootstrapping</li> <li>Model Comparison</li> </ol>	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

Introduction to An Introduction to Computational Data Analysis for Biology

## 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?

 A programming language uniquely developed for statistical analysis

Introduction to An Introduction to Computational Data Analysis for Biology

## What is R Studio?



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

Introduction to An Introduction to Computational Data Analysis for Biology	Introduction to An Introduction to Computational Data Analysis for Biology
Let's Fire It Up!	What do you see?
Open R-Studio. Don't have it? Download it from http://rstudio.org	Image: Notes of the second control of the second contententent control of the second control of the second co
The Console and Math	Everything is an Object
1 + 1 ## [1] 2	a.number <- 1 + 1 a.number ## [1] 2

Introduction to An Introduction to Computational Data Analysis for Biology	Introduction to An Introduction to Computational Data Analysis for Biology
Note: Comment Your Code as You Write with $\#$	Functions Work on Objects
The text after # is not evaluated. # This is going to be the number two a.number <- 1 + 1 ###### # You can get reative with comments to separate code # blocks and write a lot, which is good practice ######	<pre>sin(a.number) ## [1] 0.9093 How to get help for a function '?'(cos) help(cos) '?'('?'("cosine function"))</pre>
Introduction to An Introduction to Computational Data Analysis for Biology	Introduction to An Introduction to Computational Data Analysis for Biology
Lots of Object Types - like Data!	Graphics are a Snap
<pre>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</pre>	plot(speed ~ dist, data = cars)
names(cars) ## [1] "speed" "dist"	<sup>1</sup> <sup>1</sup> <del>[0 0 − 1 − 1 − 1</del> ] 0 20 40 60 60 100 120 dist

Look at ?plot to see other arguments to change appearance

### Installing Packages

its dependencies now.

	to Barrie Instan				
1 mmm	Check for the second second				
I HUMAN IL.	Warning Carmed a		in a	Date of Leasers	
	Bel.,		Man at setting an	and one of the local division of the local d	
			17	ingente service	
			All Annotation	*	
			All, Colta, Salasari) All (mail)		
			Party set as my lot any data of the set of t	• E.B., stranged - 17 ard, even data and part, "With Tand Lot", "we with even data and part," With Tand Lot", "we	Children and the second s
			In lend, oil spinned, your former, here	1980 Selfs amondes a self-self, hard,	or'), loui - 6.8, gin tar - 'arm', with m - 1. Mr. of a
			Contract Contraction	the state of the second s	A second manufacture in the second se
II. Mitshed		100	The Res Print St.	per sere r s, me mpril s 5, in 14 s 5, in	and 11, and 10, 1, 1, 10 (10)
Family Cold (BUR) (Cold (B)			Charlie & ten / burn	8 mil	
Gegrige (E) 201 No. 1 Pandettes for Eastern	of beauting		Care Parts In		to what
A In free sphere are seen and MERCHART	what's		C C LANCE		
No an annual to refer this is sale artists for the second of the formation of the second of the seco	and to		C di sans		
Betred language suggest hat rowing in an ing	tisk basis		D G nate		
The second prior prior will be a second to the second seco	All of loss				121414 1014.000.0004
has 'must'r for any press, 'weat'r for an in	an heip, or				
"Ma starty" for an UK simple' sharfes to: "Spr "K" to get 8.	ww.				
Defenses tasks from of Atming					
tion to An Introduct alling Pac	ion to Computation:	il Data Anal	ysis for Biology	/	
ion to An Introduct alling Pac	ion to Computation:	il Data Anal	ysis for Biology	1	
tion to An Introduct alling Pac	ion to Computation: ckages	il Data Anal	ysis for Biology	/	
alling Pac	ion to Computation:	il Data Anal	ysis for Biology	(	
tion to An Introduct alling Pac	ion to Computation: ckages	il Data Anal	ysis for Biology	/	
alling Pac	ion to Computationa	il Data Anal	ysis for Biology	1	
alling Pac	ion to Computationa	il Data Anal	ysis for Biology	,	
alling Pac	ion to Computationa	il Data Anal	ysis for Biology		
alling Pac	ion to Computation:	il Data Anal	ysis for Biology	,	
ion to An Introduct	ion to Computation	Il Data Anal	ysis for Biology		
ion to An Introduct alling Pac	ion to Computationa	il Data Anal	ysis for Biology	command li	78
ion to An Introduct alling Pac	ion to Computationa ckages	I Data Anal	ysis for Biology	, command li	ine.
ion to An Introduct alling Pac You can also	ion to Computationa Chages	il Data Anal	ysis for Biology rom the o	, command li	'ne.
alling Pac	ion to Computation ckages o install pack	I Data Anal	ysis for Biology	, command li	ine.
ion to An Introduct alling Pac You can also install.pack	ion to Computations	al Data Anal kages fr	ysis for Biology rom the o	command li	ine.
alling Pac rou can also nstall.pack	ion to Computations ckages o install pack ages ("ggplo	I Data Anal xages fr	rom the o	command li	ine. case.edu/",
alling Pac	ion to Computation: ckages o install pack ages ("ggplot cies = TRUE;	al Data Anal kages fr	rom the o	command li	ine. case.edu/*,
fou can also nstall.pack dependen	ion to Computations ckages o install pack ages ("ggplot cies = TRUE)	il Data Anal kages fr 2 <sup>2</sup> , rej	rom the o	command li	ine. case.edu/*,
alling Pac fou can also dependen	ion to Computations ckages o install pack ages (*ggplot cies = TRUE)	al Data Anal kages fr	rom the o	command li	ine. case.edu/",
alling Pac	chages b install pack ages ("ggplot cies = TRUE)	al Data Anal kages fr 2", rej	rom the o	command li	ine. case.edu/*,
ion to An Introduct alling Pace You can also dependen Using one of	ion to Computational chages o install pack ages ("ggplot cies = TRUE) f the above	kages fr 2", rej method	ysis for Biology from the o	command li http://cran.	ine. case.edu/", ze grplot2 and

#### Installing Packages

		Rhafe			
Comments of the second s		Andrews Manual			
Contraction in day	and in class of	which it has strengthere	and diverse		
		100			
		10,70	January, of A contraction		
		100	and a second		
		a second	in the second se		
		la m	and the second s		
		100	1981		
		Andrea			
		all, forma (alphane)			
		(Marchan 1			
		ray interior, the int, type	1.6.4. Subjects 1.75	the local state of the second second	and the second se
		anter war of , seprenduke to	stationerskill, tagent lat - maniaclasticitum	Chainerith, man - Wat of Marty', pit + 1.	n.men.d., big -
		in les A, of + place), you			
		"Industry, with a description		start is made a \$10, and take a particular a	
and the second se	(marketer)	to service a service part as	<ol> <li>Andre 1 L. Ten, Separat 1 T. Gall, M. C. L. H.</li> </ol>	RATE OF BETTER AND A REPORT	LILA, M. April 1.
	hotter				
	Annual Contract of the		a second s		
¥ spruise 5 (0.2 (000) 40 (0)					
THE LOOP ALL ALL	Participa departe	a prefit age and a model.		14	marine a
					No. 24 (011), 1 (0 Her
A 12 Area and and and and and and and and an and an and an and	Insul account			25.7.98	Bec 74, 2114, 811, 74
To an arrive to refer to a lot offer and then.	(1000)	arte & fundamente ferrenzen 2.1.8 Marco 10		1998 10400	Apr 4, 1011, 412174
"gas 'timesely or 'timesely' for dustriant destin-				10/2148	Her   F 100 F 11 h we
Month language search had particle to an institute language					
It is a satisficant a proper will keep partition of					
"schedule")" on the forming in the second or additional		C. C. Berne			
And insult to one have been been to be a long him have as		The second		100 C 100	And in case of the loss of the
"wip starts?" for or 400, proper interface to help		CO 13 hour water of		10710-044	No. 4 COM L DOM: N
for 'e)' to ant t.		C 10 hotel, cas			
Declagor haded from of Manual		O C gran paters marger had		0.4.48	Aug. 1, 1911, 1918, 191
		O produces making			Sec. 311, 1974
		C C market, est-da		1.4.49	Res. 5, 10010, -12-10, 000
		O R webbererstyle			
		0 Gardent		10.1148	
		C C and at		1.4 mb	0410, 100, 111, M
		0 C geographics		107.008	No. 1, 1011, 1101 Mg
		10 L1 mm		100.1 08	Res 1, 2010, 13, 18, 99
		the states			
		Contraction of the second seco		1007.0.000	March, 1910, 1111, March 1910
		and the second			
		U C Are			
		U D Nove			
			1000 P111P (P1)		

#### Introduction to An Introduction to Computational Data Analysis for Biology

#### 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



#### Next time

Data Management!

Introduction to An Introduction to Computational Data Analysis for Biology

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

# Questions?