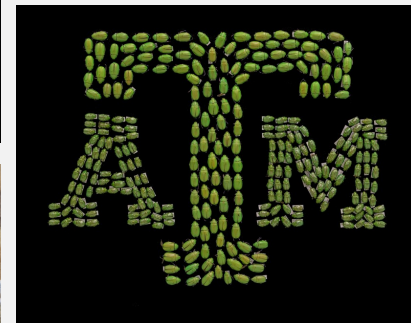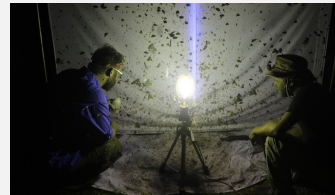# Experimental Design
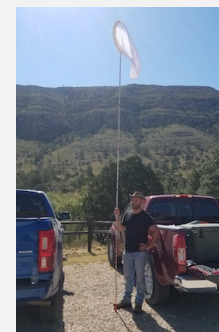## Biology 683

Lecture 1

Heath Blackmon
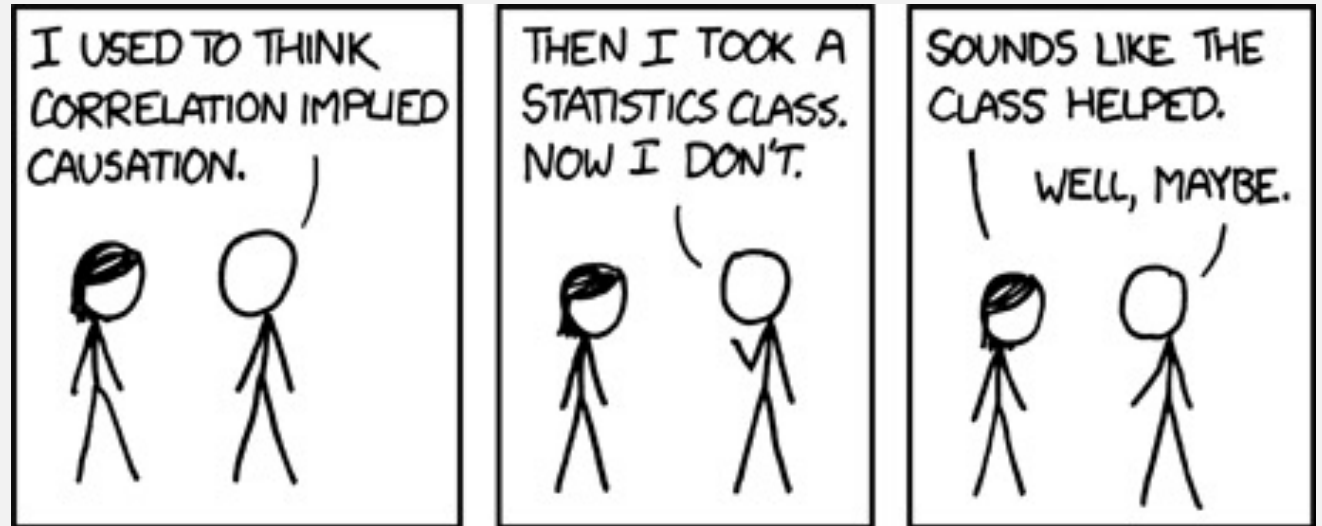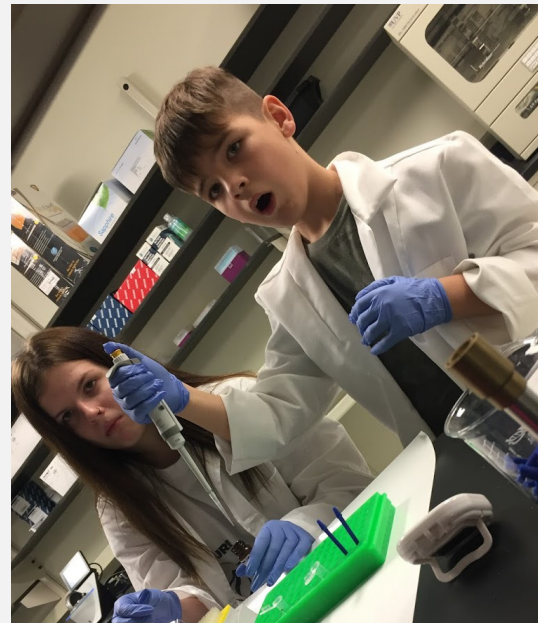
Me

I study the **evolution of traits**. Especially genomic traits. I use a variety of methods next-gen sequencing, experimental evolution, phylogenetic comparative methods, and theoretical approaches. I don't work with a single taxa; we have projects involving fish, mammals, reptiles, amphibians, insects, and bacteria all ongoing in my lab.

You all

Fill out index cards for me

My Objectives

- *Help you build an intuitive understanding of statistics*

- *Help you develop the confidence to think about the characteristics of the data that you will be collecting in your research and how you might analyze it.*

- *Get you comfortable with the idea of coding in R*

- *Help you develop the skills to handle datasets in R*

- *Help you develop the skills to build informative, honest, and intuitive data visualizations in R*

- ***Make you a more productive and successful scientist!***

My View on Graduate Courses

- Graduate courses should open the door for you to become an expert in a field.

- However, you shouldn't need to become an expert to do well. Not every class is super central to your research.

- Therefore, I'm going to expose you to a lot of material and hope to really challenge you with some of the problems we solve. However, you can earn an A if you put forth an appropriate and reasonable effort.

- I hope that some of you embrace the skills that I am teaching you, as I believe they can be crucial to becoming an outstanding scientist.

Pedagogy

Class Size and Student Achievement:
Research-Based Policy Alternatives

Allan Odden
University of Southern California

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Active learning increases student performance in science, engineering, and mathematics

Scott Freeman, Sarah L. Eddy, Miles McDonough, +3, and Mary Pat Wenderoth   Authors Info & Affiliations

RESEARCH ARTICLE | APPLIED PHYSICAL SCIENCES

Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom

Louis Deslauriers, Logan S. McCarty, Kelly Miller, +1, and Greg Kestin   Authors Info & Affiliations

What I will offer

- In class (MW 4-5:15) – lecture, discussion, questions and answers, live coding

- Monday Code Nights (MW 5:30-6:30) – week before tests

- Saturday Code Days (9-12) donuts provided: Fill out form with weekends that you could attend.

- Lots of enthusiasm!

Today

- Pedagological approach
- Syllabus / website / calendar
- Why I think you need this class
- R

My opinions

Misuse of statistics is unethical as a scientist

My opinions

Misuse **or ignorance** of statistics is unethical as a scientist

Poor training and maleficence are both responsible for failures

Statistical literacy in the general public is essential and lacking

Do your part: learn science of important topics and help friends and family understand them! **This includes the statistical analysis and how we should let them inform our belief!**

Reproducibility crisis

- Started in the social sciences but some problems are widespread

- p...

- f...

- small sample sizes

- p-hacking

- unethical researchers/developers

Discuss possible causes of the reprodu...sis

	All	Since 2019
Citations	1634	1305
h-index	20	20
i10-index	27	26
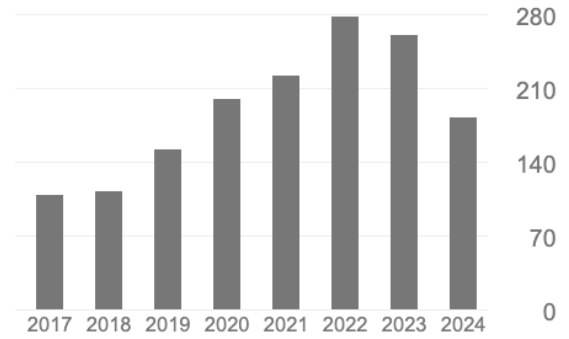
Jonathan Pruitt
Canada 150 Research Chair

Amy Cuddy
TED Talk 72.2 Million views
(2nd most popular TED Talk)

- Study preregistration ([COS](#))

- Discuss What are possible solutions to the reproducibility crisis?

- Systemic change - unlikely

Why do biologists need statistics

- We want to learn about the world often by testing hypotheses.

- To test a hypothesis we have to design an experiment

- Not all experiments have a traditional control and experimental treatment and this isn't always how we want to test a hypothesis

- It is quite possible to design a study or collect data that cannot answer the questions that we have

- This leads to poor manuscripts and can lead to bad practices like p-hacking

Experimental Design

To design an experiment you need to understand how the data will be analyzed statistically.

1. How can you sample the population in which you are interested?

2. What tests are appropriate for your data?

3. What biases must be controlled for?

4. What sample size will be necessary?

Why not just collaborate with a statistician

1. In some cases this is a great option, but you have to understand enough to communicate.
2. If you publish a study you are responsible for its validity.
3. For most experiments simple methods suffice.
4. In many fields of biology there are sets of statistical tests that are expected for certain types of data.
5. For all of these reasons statistical analysis **needs to involve people who understand the biological problem and the field of study.**

My stats philosophy

- Statistics is just another tool

- My responsibility as a scientists is to report the truth as accurately as possible and statistics help me in this regard

- We may NEED statistics to discern patterns in our data

- You need to understand where the signal that makes for a significant test comes from. Visualizing your data in the right way can do this!

Why am I teaching this class?

Evoldir Postdoc Adds

December 1, 2017 – January 15, 2018

74%
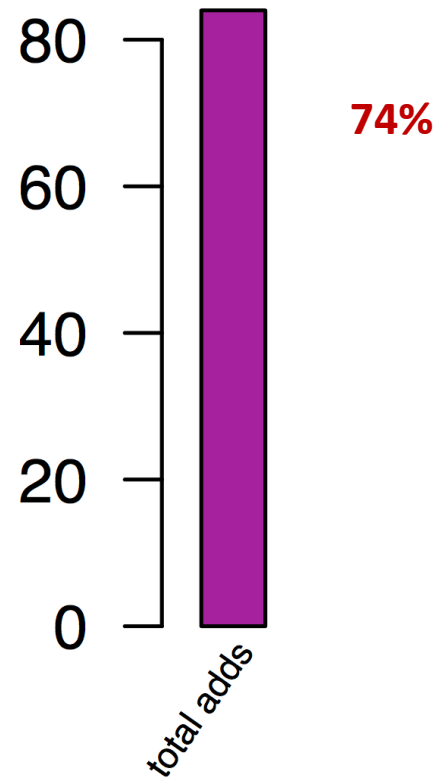
What is R

- R is an open and free statistical programming language that focuses on stats and graphics

- It works very similarly on all major operating systems

- It's also a full-fledged high level programming language (similar to Python)

- Very popular in industry so looks great on a CV.

Why use R

1. Many statistical approaches have been implemented in the R environment.

2. Because it's open source, there are no proprietary secrets, as might be hiding in commercially available statistical packages.

3. Any program written in R will have access to all of R's tools for statistics and graphing.

4. New methods of analysis are being implemented in R by the scientists developing the methods.

Why use R

5. If you use R you can include a script with your manuscript [example]

     Reproducibility / Open science

     Reviewing

     Revising

6. Many methods (mixed models, quantitative genetics, etc.) are only available in R.

7. PLOTTING

8. Once you've learned one language you can learn others more easily.

Downsides of R

- Learning curve
- Anyone can make a package - so there is some junk out there
- Memory issues
- No language lasts forever and no language can do everything
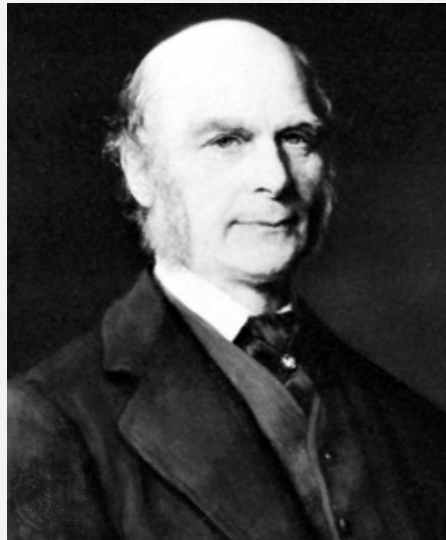 - ○ Python
  - ○ Awk
  - ○ Julia

The Origin of Statistics

Much of modern statistics was an offshoot of genetics and evolution

K. PEARSON
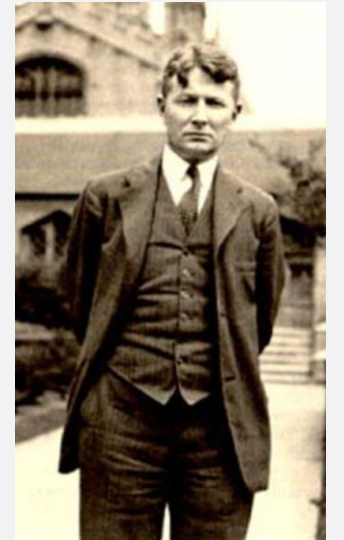1857-1936
CORRELATION

F. GALTON
1822-1911
REGRESSION

R. FISHER
1890-1962
ANOVA

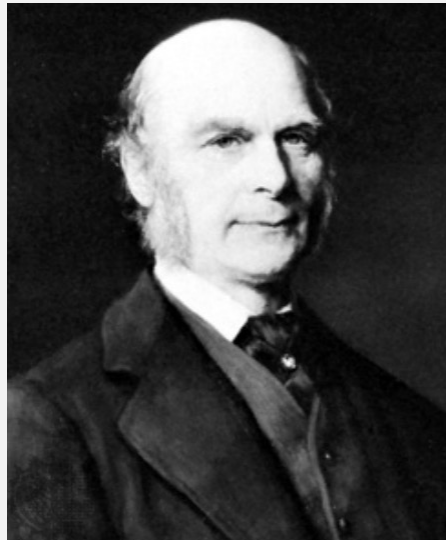
S. WRIGHT
1889-1988
PATH ANALYSIS

1900 rediscovery of Mendel's work was motivating problem.

The Origin of Statistics

K. PEARSON
1857-1936
CORRELATION

F. GALTON
1822-1911
REGRESSION

R. FISHER
1890-1962
ANOVA

The disgraceful history of biostatistics

- Much of statistics was developed with the idea of showing that we could measure, scientifically analyze, and improve the "quality" of humans.

- The majority of geneticists and statisticians in the early 1900s were proponents of eugenics.

- What are the problems with this scientific/ethical
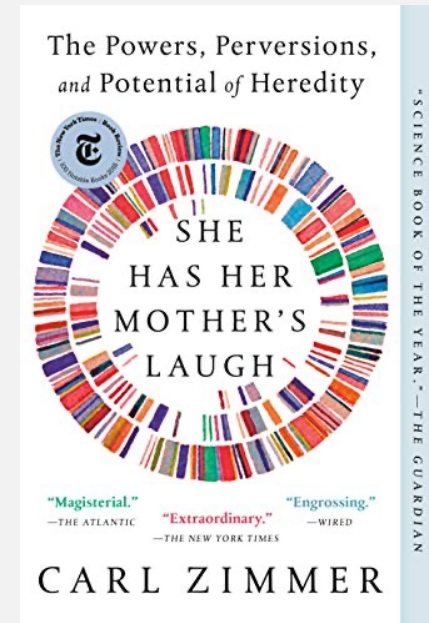
For Future Classes

1. Install R and Rstudio on the computer you will use this semester
2. **See me if you have problems installing it.**

You can bring your laptop to class to follow along on coding that I do in front of you but this is not a requirement. Our room has insufficient plugs so charge ahead of time. **I expect you to practice outside of class and come to practice sessions if you have problems.**

Heath Blackmon
BSBW 425
Blackmon@tamu.edu

Installing R and RStudio

Installing R
1. Go to the R homepage and click download R.
2. Pick a mirror that is in Texas or at least in the United States.
3. Select the correct version for your system and follow the prompts.

Installing Rstudio
1. Go to the RStudio homepage and click on the download link below the free version of RStudio Desktop.
2. Select the correct version for your system and follow the prompts.

How you will be learning
1. I will code live in front of you, I will have some days times that we set aside extra time for help. HOMEWORK!

Basics of R

1. Demo R

Data structures
- vector
- matrix
- dataframe
- list

Data types
- numeric
- character
- logical
- factor

Control elements
- for
- if
- while

Common functions
- c
- matrix
- list
- sum
- mean
- sd
- sqrt
- abs
- paste
- rnorm
- rbinom
- rexp
- sample
- rep
- data
- Help
- which

Basic base R plotting functions
- hist
- plot
- density
- abline
- lines

Operators
- <-
- ==
- >
- <
- %in%
- {
- [
- + - * / ^ %%

Practice

```
install.packages("swirl")
library("swirl")
swirl()

Complete two lessons of your choice
```