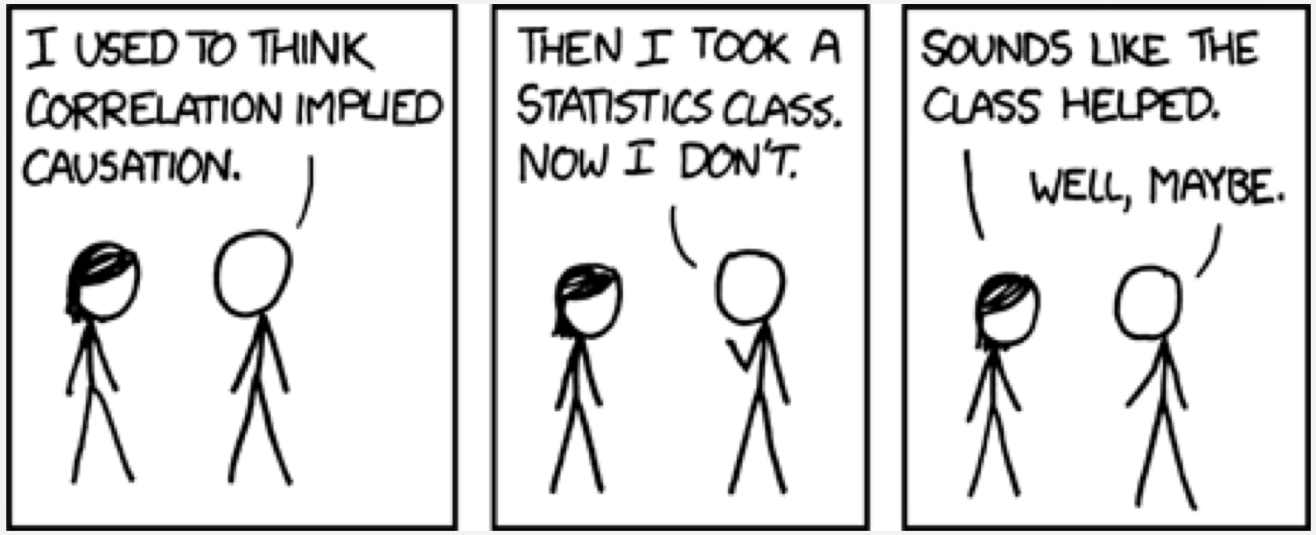
Experimental Design

Biology 683

Lecture 1

Heath Blackmon



Today

- Introductions
  1. Name
  2. Lab
  3. Project / Data
- Syllabus / website
- Big problems in stats (outside world / within academia)
- Why you need this class
- Prep for Thursday

My Objectives

- *Help you build an intuitive understanding of statistics*
- *Get you comfortable with the idea of coding in R*
- *Help you develop the skills to build informative, honest, and intuitive data visualizations in R*
- *Help you develop the skills to handle datasets in R*
- *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.*

The public impression of statistics

- *There are three kinds of lies: lies, damned lies, and statistics*
- *You can make statistics say anything*
- *Statistics are no substitute for good judgement*

My opinions

Misuse or ignorance of statistics is unethical as a scientist

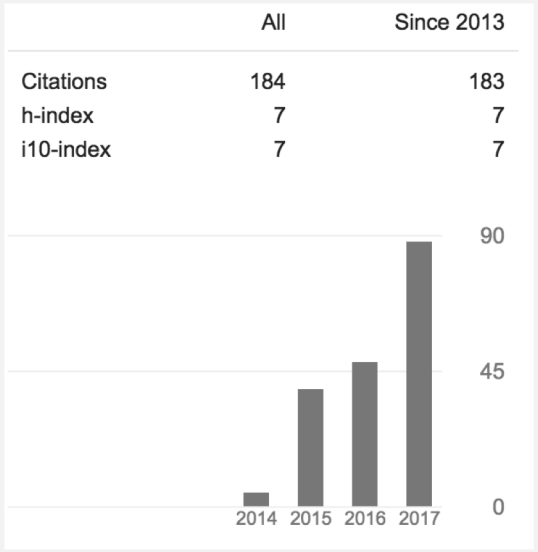
Poor training and maleficence are both responsible for failures

Statistical literacy in the general public is essential

Do your part: learn science of important topics and help friends and family understand them! This includes the statistical analysis

Reproducibility crisis

- Started in the social sciences but some problems are widespread
- pressure to publish
- file drawer problem
- small sample sizes
- p-hacking
- unethical researchers



Amy Cuddy
TED Talk 47 Million views
(2nd most populat TED Talk)

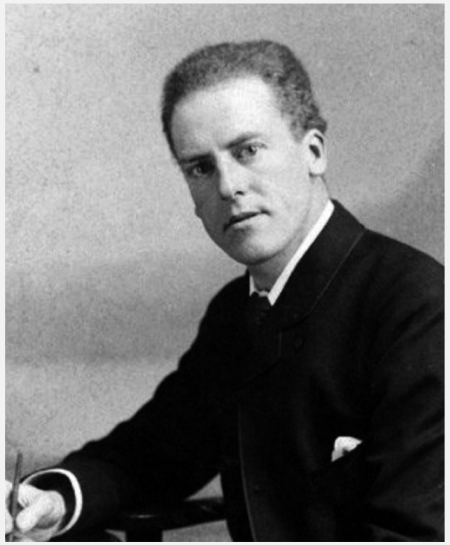
Solutions

- Study preregistration
- PeerJ / PLOS ONE
- Preprint Servers
- Altimetrics
- Systemic change - unlikely

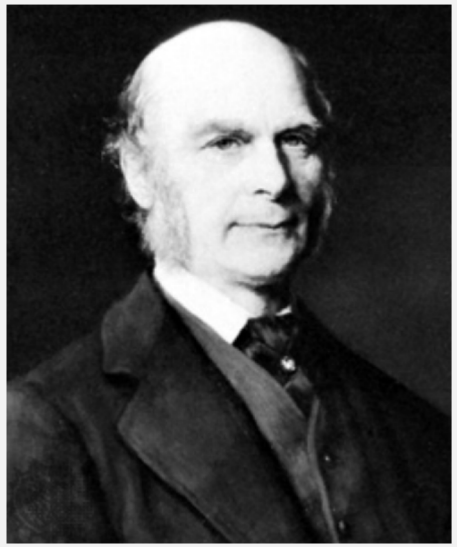
The Origin of Statistics

In many ways modern statistics was an offshoot of evolutionary biology (1900 rediscovery of Mendel's work was motivating problem).

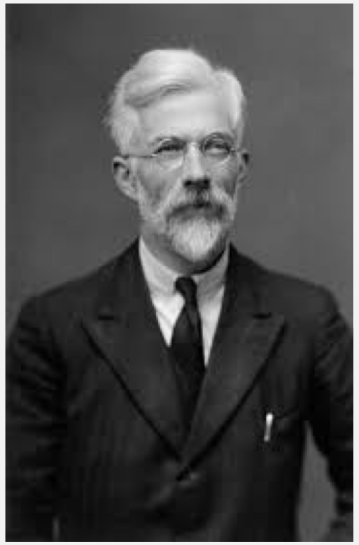
K. PEARSON
1857-1936
CORRELATION



F. GALTON
1822-1911
REGRESSION



R. FISHER
1890-1962
ANOVA



S. WRIGHT
1889-1988
PATH ANALYSIS

Why do biologists need statistics

- We want to test 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 – or mastering out

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?

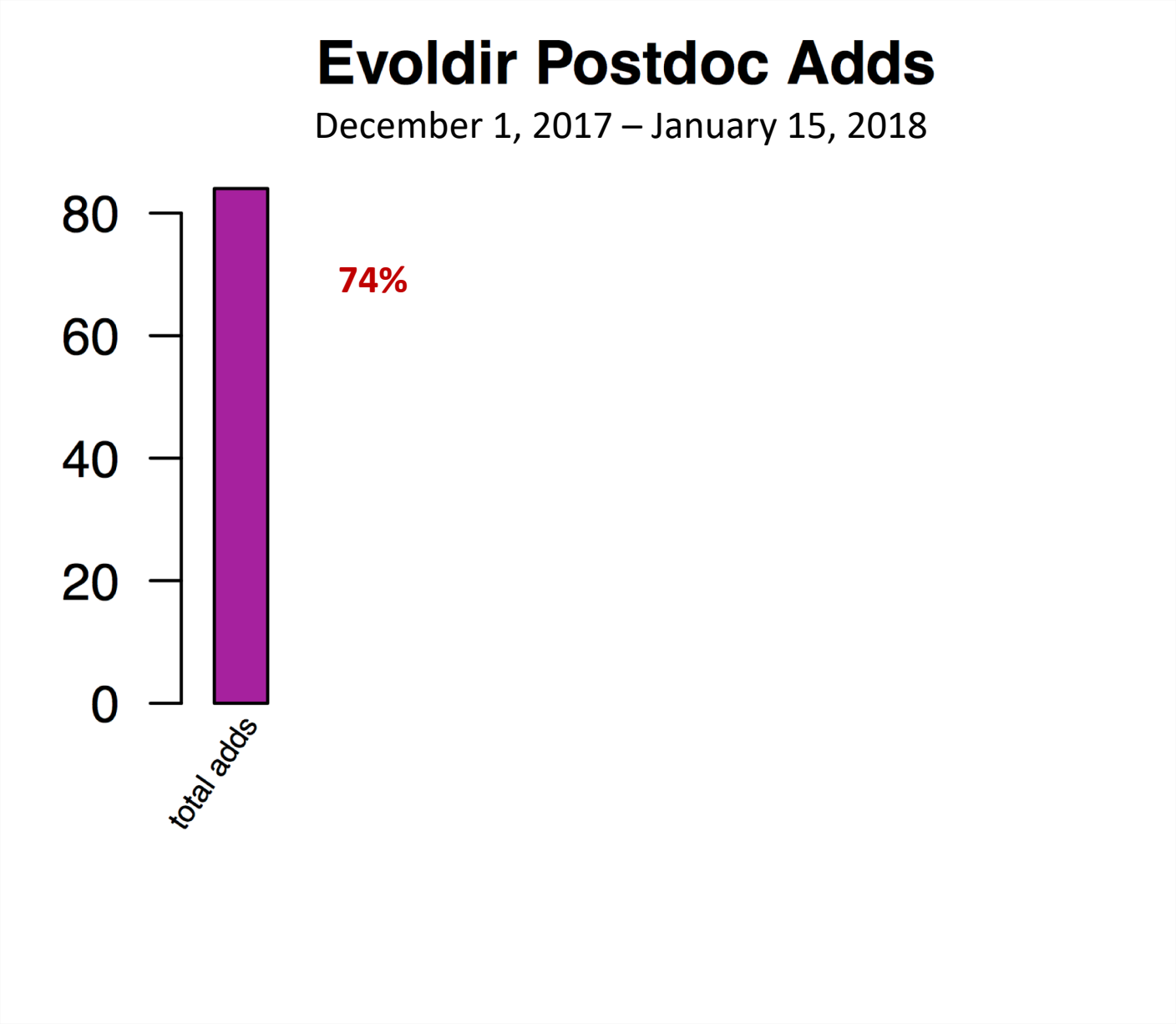
Why not just collaborate with a statistician

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**

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?



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)
- *FYI: 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
   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

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.

For Thursday

1. Do homework 1.
2. Install R and Rstudio on a laptop
3. Come and see me **BEFORE** class on Thursday if you run into problems
4. Read chapters 1 and 2 of WS – good supplemental readings too!

Bring laptop to class every day from here on out! Bring a charger if you are not 100% positive that your battery will last.

Heath Blackmon
BSBW 309A
coleoguy@gmail.com