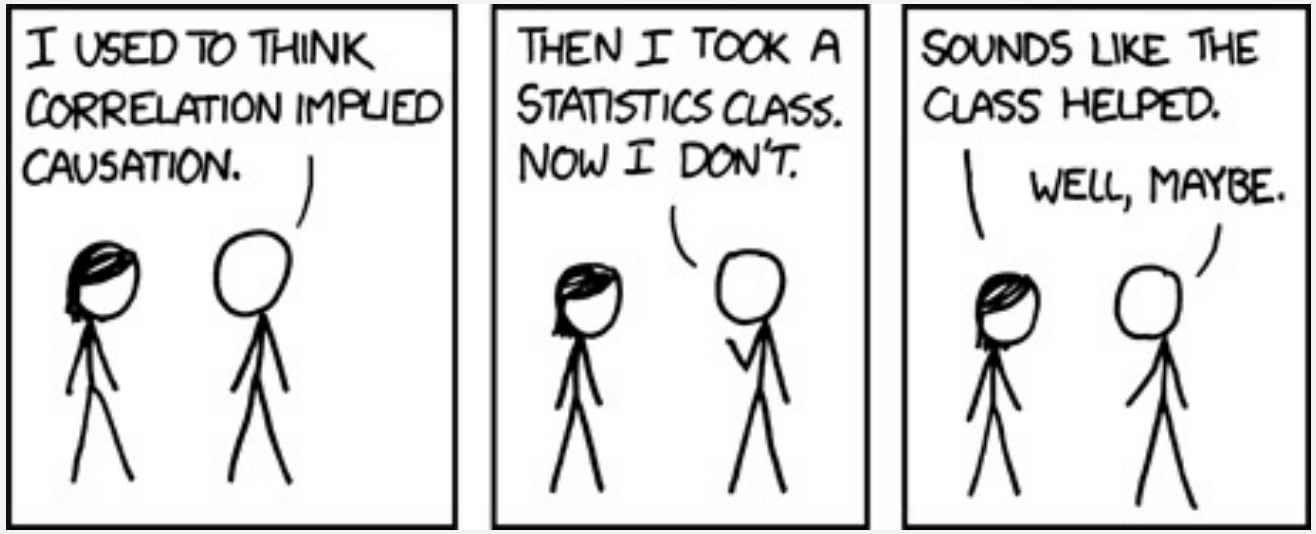
Experimental Design

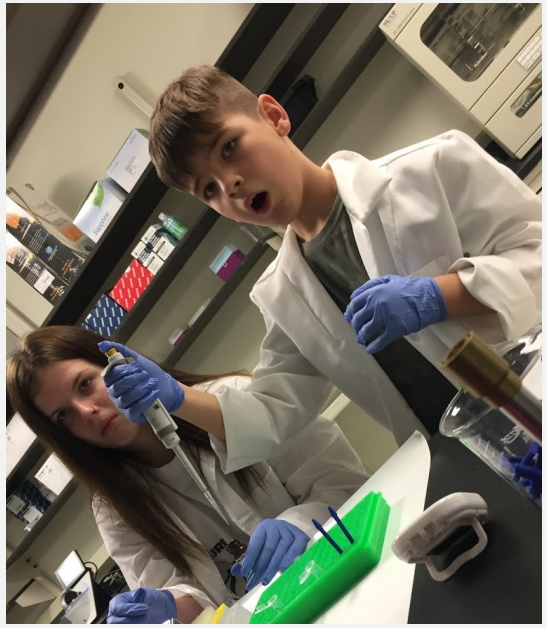
Biology 683

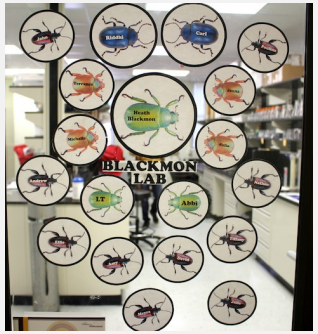
Lecture 1

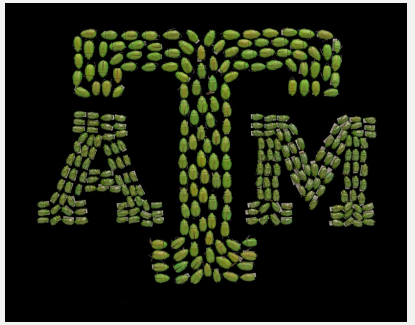
Heath Blackmon



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.

Me

2010-2015 Graduated PhD from UT Arlington
2015-2017 PostDoc UM
2017-present TAMU (37 Papers since starting my lab)

Assoc Dept Head Biology
Chair EEB IDP

NIH Funded

- Sexual antagonism
- Genome structure
- Epistasis

Passionate about PhD training and how we can make it better

Blackmon Lab: 5 Grad students, 12 Undergrads, Postbac, Postdoc, Research Scientist

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!***

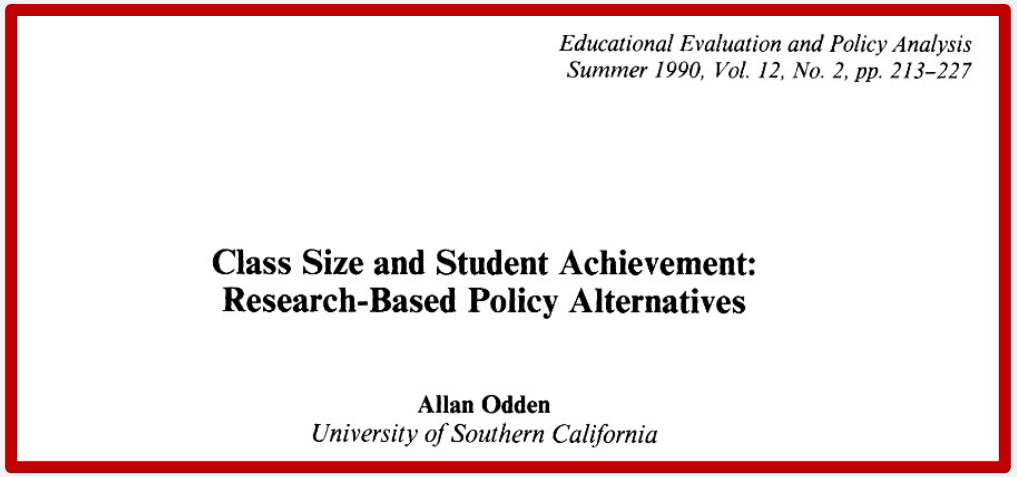
My View on Graduate Courses

- 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.

Today

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

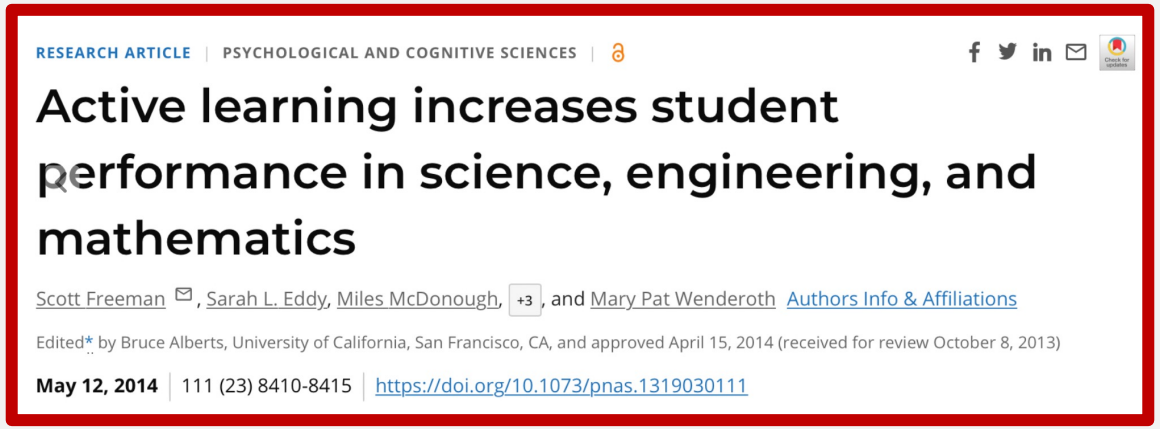
Pedagogy



Educational Evaluation and Policy Analysis
Summer 1990, Vol. 12, No. 2, pp. 213–227

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

Edited* by Bruce Alberts, University of California, San Francisco, CA, and approved April 15, 2014 (received for review October 8, 2013)

May 12, 2014 | 111 (23) 8410-8415 | https://doi.org/10.1073/pnas.1319030111

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

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved August 13, 2019 (received for review December 24, 2018)

September 4, 2019 | 116 (39) 19251-19257 | https://doi.org/10.1073/pnas.1821936116

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 survey with weekends that you could attend.
- Lots of enthusiasm!

Today

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

Today

- Pedagogical 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

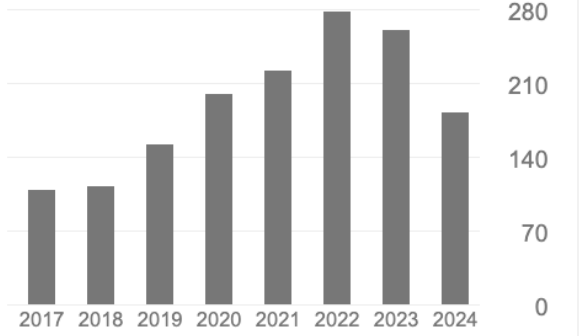
Do your part: learn science of important topics and help friends and family understand them! **This includes the statistical analysis and how we should let them inform our belief!**

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 reproducu 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)

Solutions

- 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?

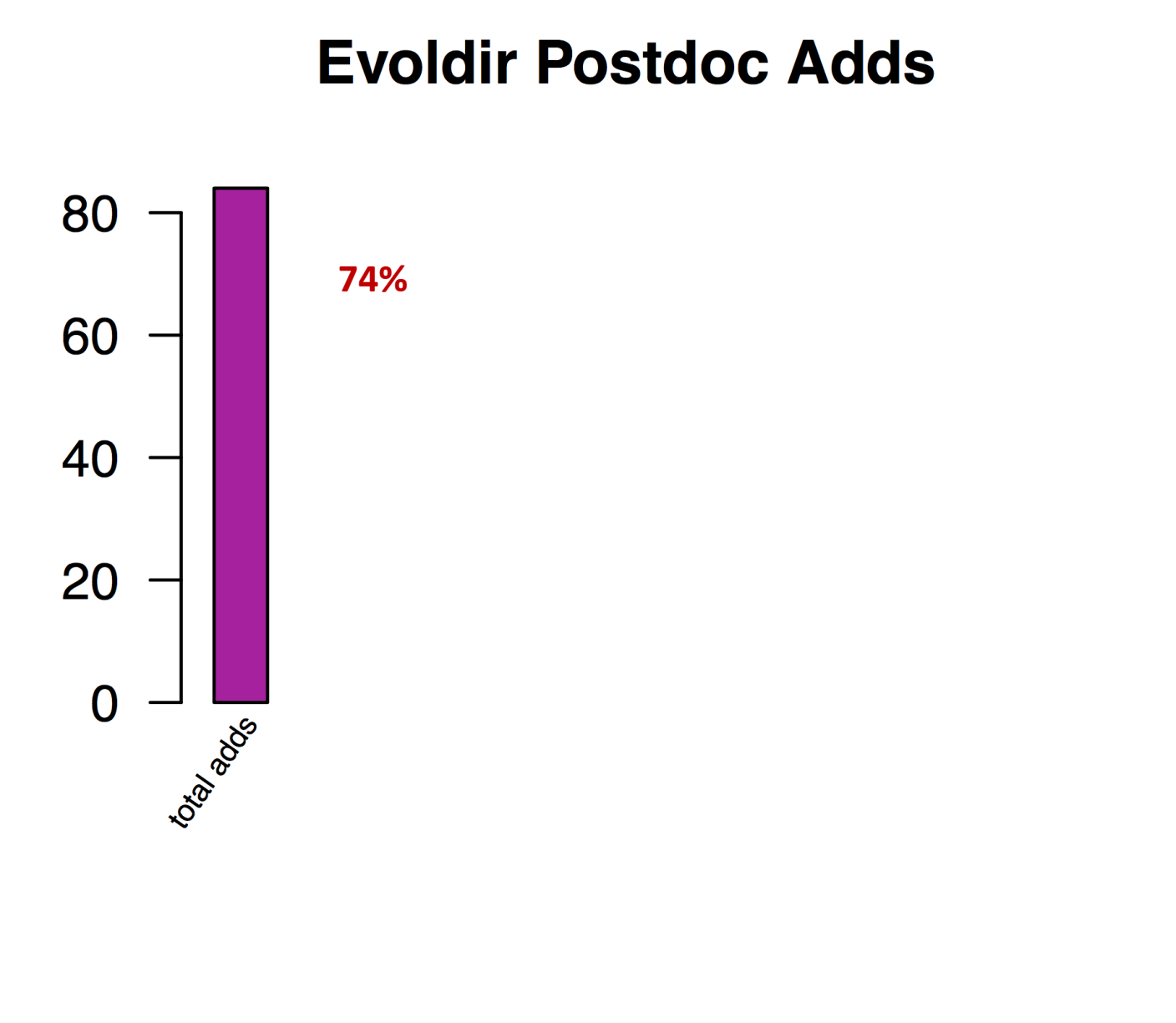
Why not just collaborate with a statistician

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 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 specific 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?



Today

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

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.

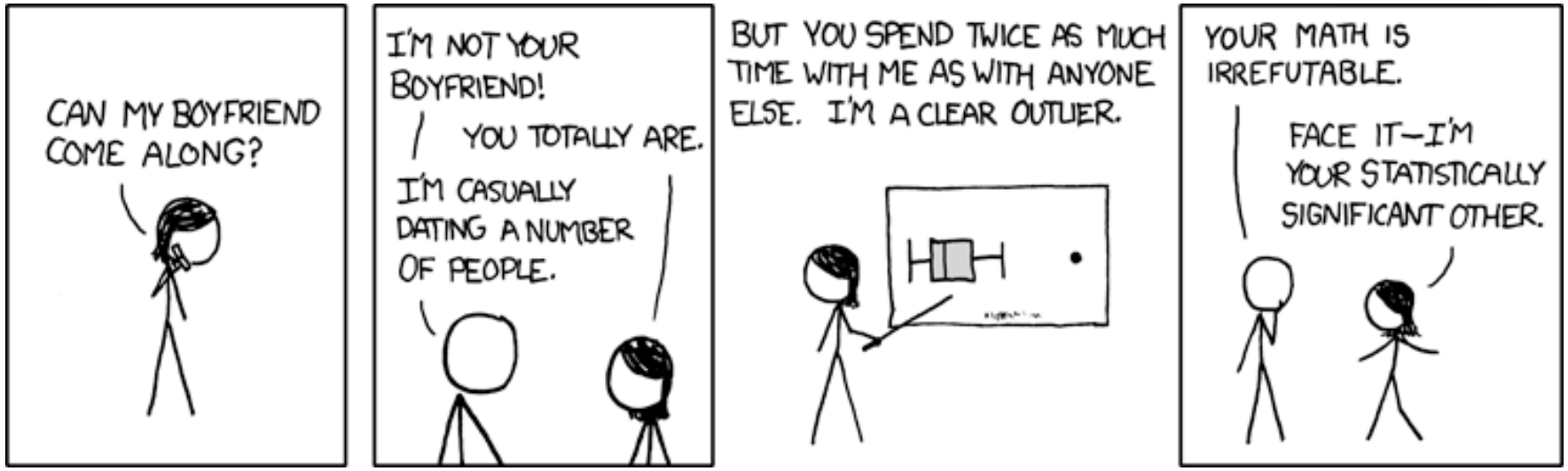
7. PLOTTING

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

Statistical Principles

Biology 683

Heath Blackmon



Example test questions:

Name 3 causes of the reproducibility crisis

Name 3 possible solutions to the reproducibility crisis

Downsides of R

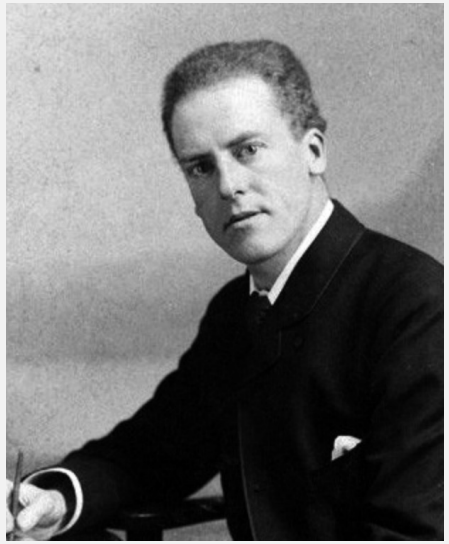
- Learning curve
- Anyone can make a package - so there is some junk out there
- Memory issues
- No language lasts forever and no language can do everything
    - Python
    - Awk
    - Julia

START HERE

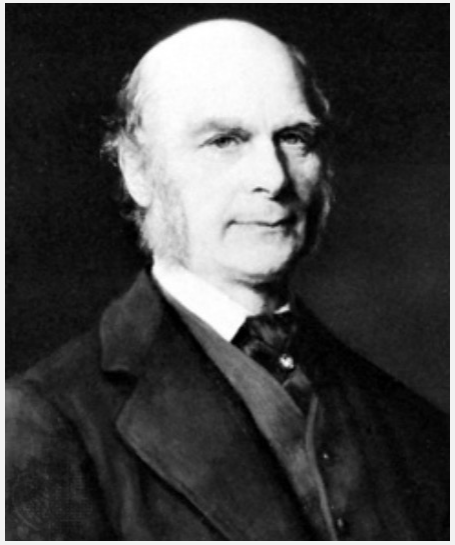
The Origin of Statistics

Much of modern statistics was an offshoot of genetics and evolution

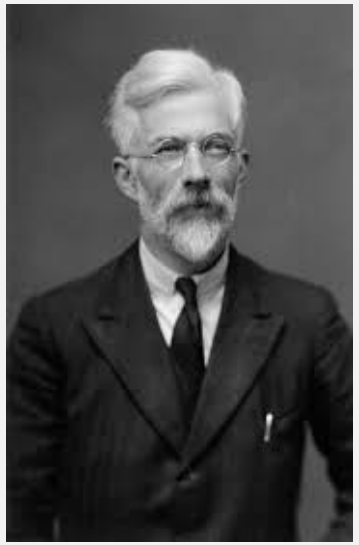
K. PEARSON
1857-1936
CORRELATION



F. GALTON
1822-1911
REGRESSION



R. FISHER
1890-1962
ANOVA

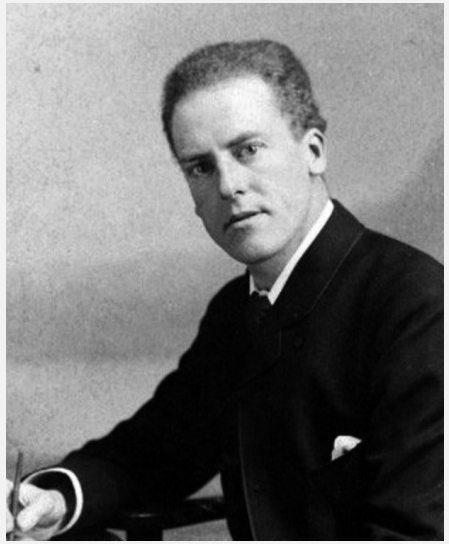


S. WRIGHT
1889-1988
PATH ANALYSIS

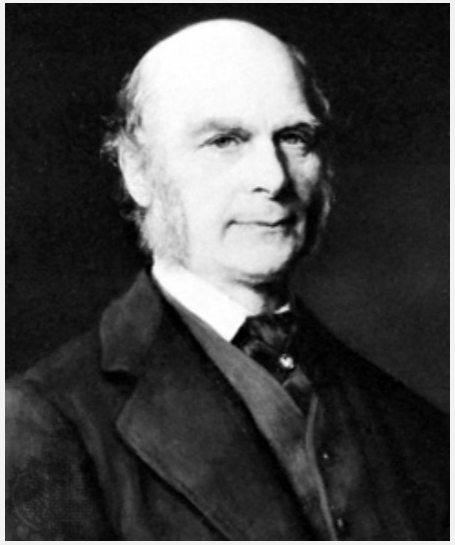
1900 rediscovery of Mendel's work was motivating problem.

The Origin of Statistics

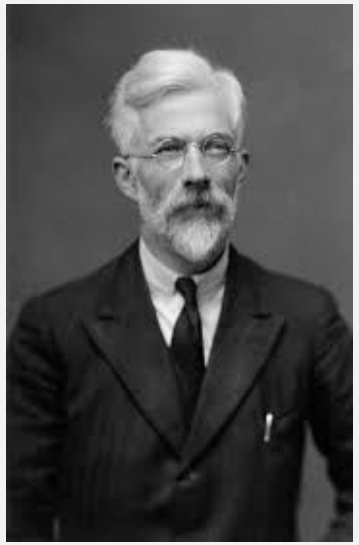
K. PEARSON
1857-1936
CORRELATION



F. GALTON
1822-1911
REGRESSION



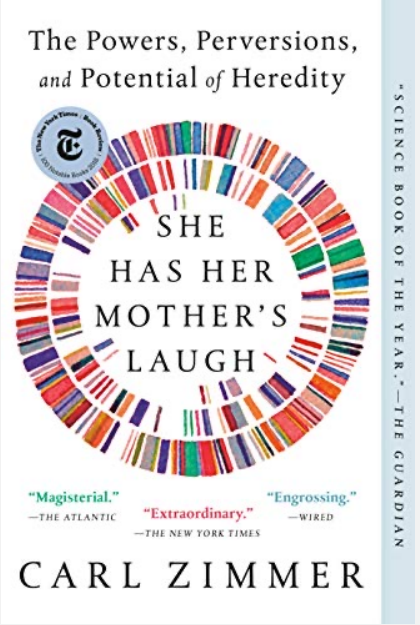
R. FISHER
1890-1962
ANOVA



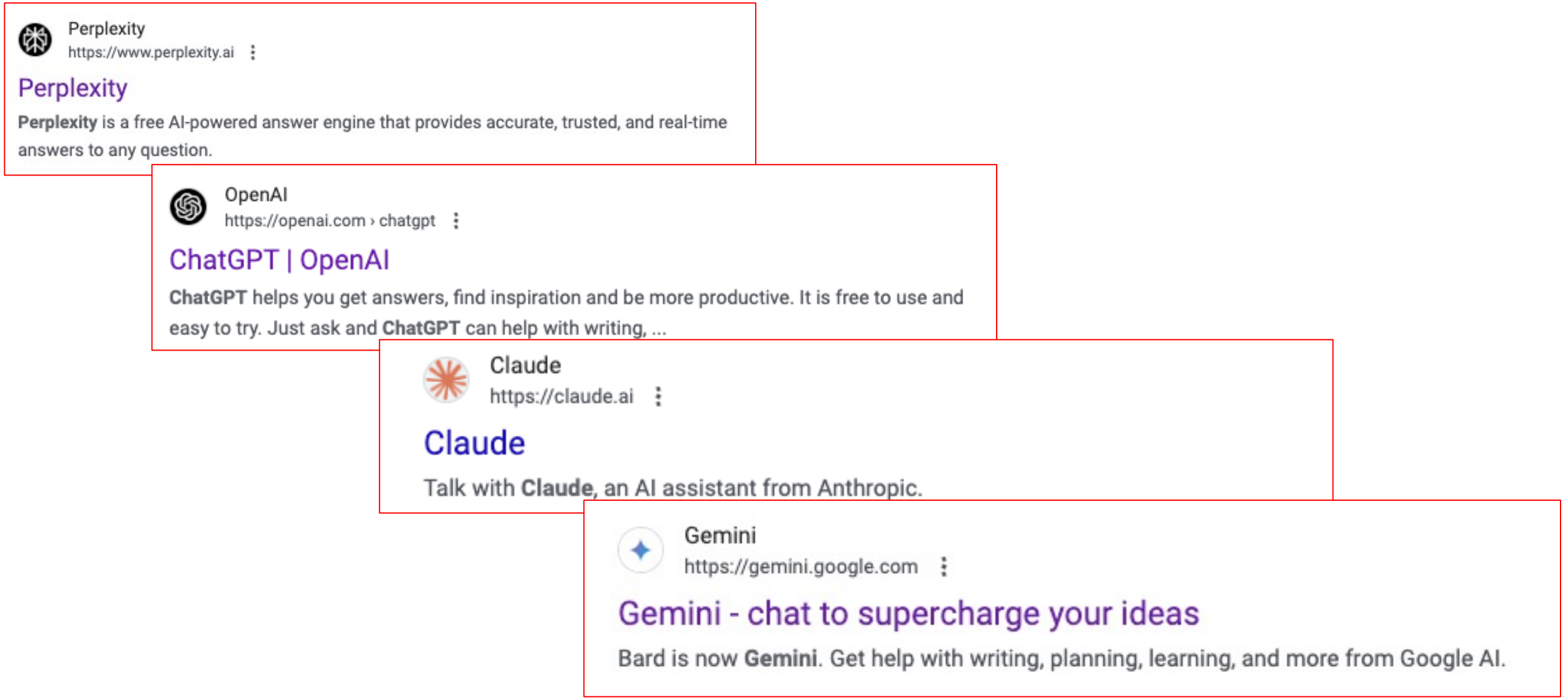
The disgraceful history of biostatistics

- 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



Landscape of generative AI



Large Language Models

- Trained on most of the words humans have stored electronically.
- The results can be shockingly good if what you need to produce is similar to many products in the training dataset.
- At their core, LLMs lack understanding of what they are doing, so they can and do produce nonsensical, incorrect responses. (Hallucination)

- Collaboration
- Understanding
- Responsibility
- Reputation

General Advice

- Do not use in any class to directly generate answers unless told to do so
- Do use it to get additional explanations of topics
- Do use it to get examples of ideas, approaches, etc.
- Do not use it to find scientific literature
- Do use it to troubleshoot improve code
- Do not use it to write your code

Rules for this class

- We will not use it for the first 1/3 or so of the class.
- I will tell you explicitly when you are allowed to use it.
- Do not use it if you are not 100% certain that I have said it is ok.

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.

This week

1. Terminology (a lot of it!)
2. P-values
3. Terminology
4. Central Limit Theorem

Data

Data terminology

Data

Variables

The characteristics that differ among individuals

Data

The measurements of variables taken for a sample of individuals

Data

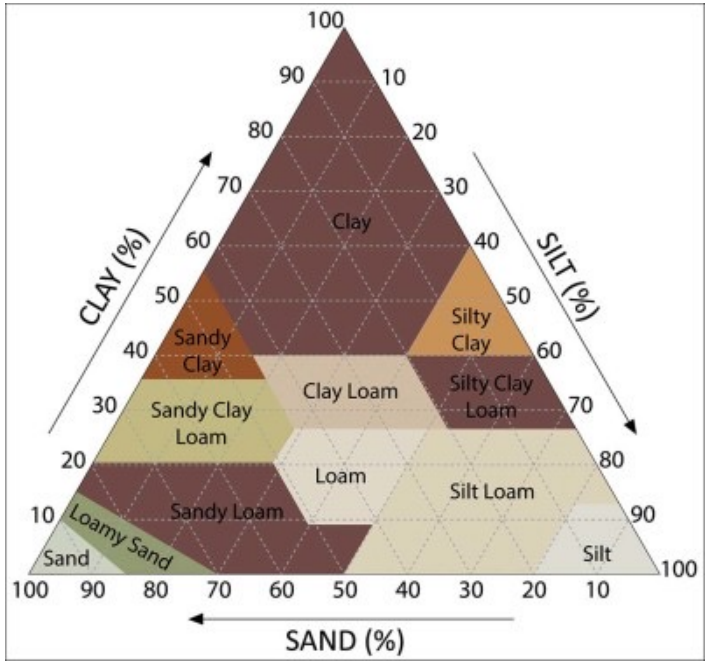
Numerical Variables
Individuals vary on a quantitative scale

Categorical Variables
Individuals are in qualitative categories

Ordinal
The categories can be ordered

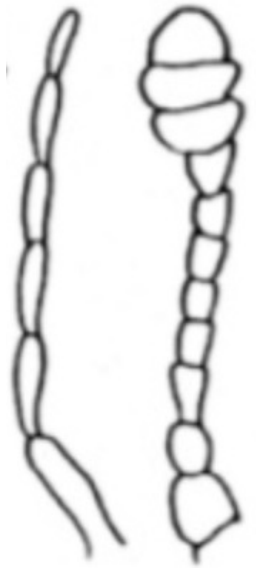
Nominal
The categories have no inherent order

Figure 1. Joiner's living histogram of student height.

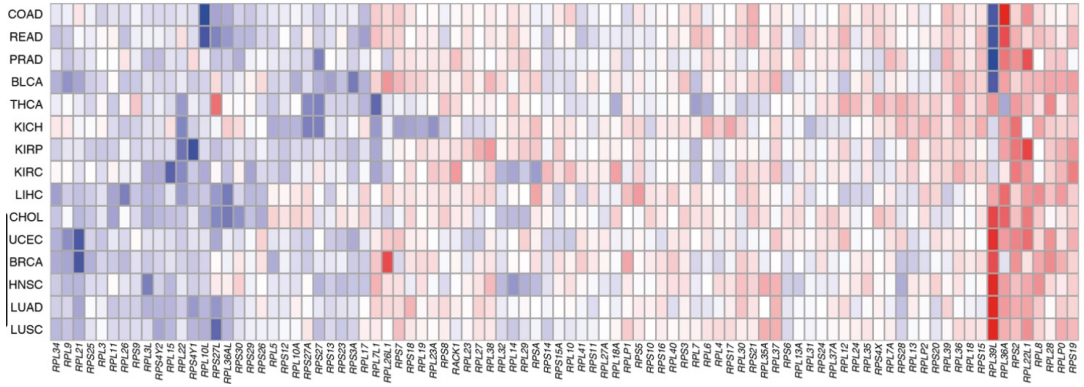


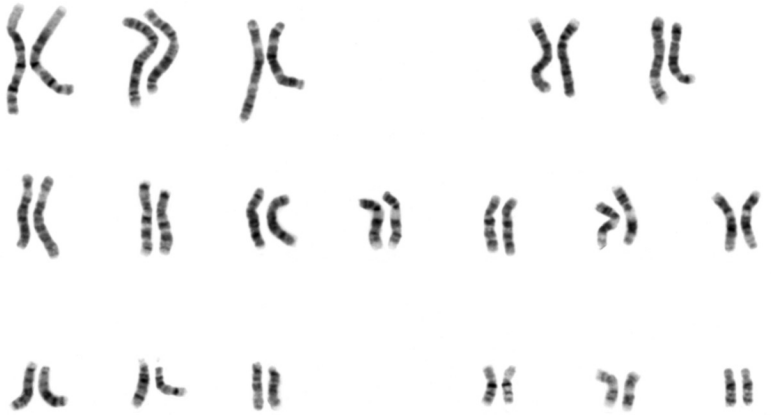
Continuous vs Discrete

Continuous variables
a variable that has an infinite number of possible values

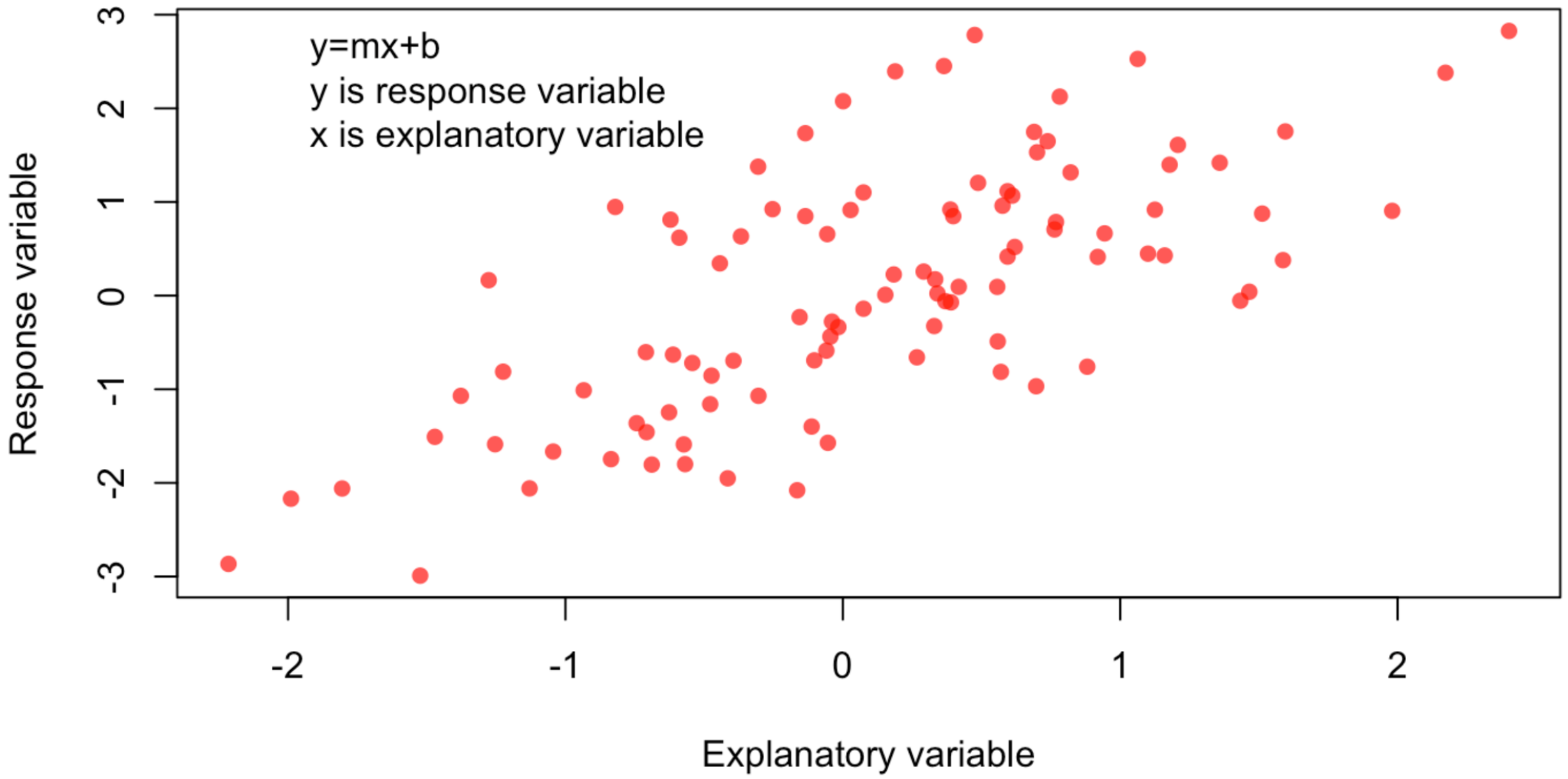


Discrete variables
a variable that has a finite number of possible values





Explanatory and Response Variables



Sampling

What is sample?

Populations and Samples

- **Populations**

  Some sort of group of something - could be anything
  - Undergraduates at Texas A&M
  - Jewel beetles in Arizona
  - Strain of flies in the lab
  - People on the titanic

- **Samples**
  - A subset of individuals drawn from a population

Going from samples to populations

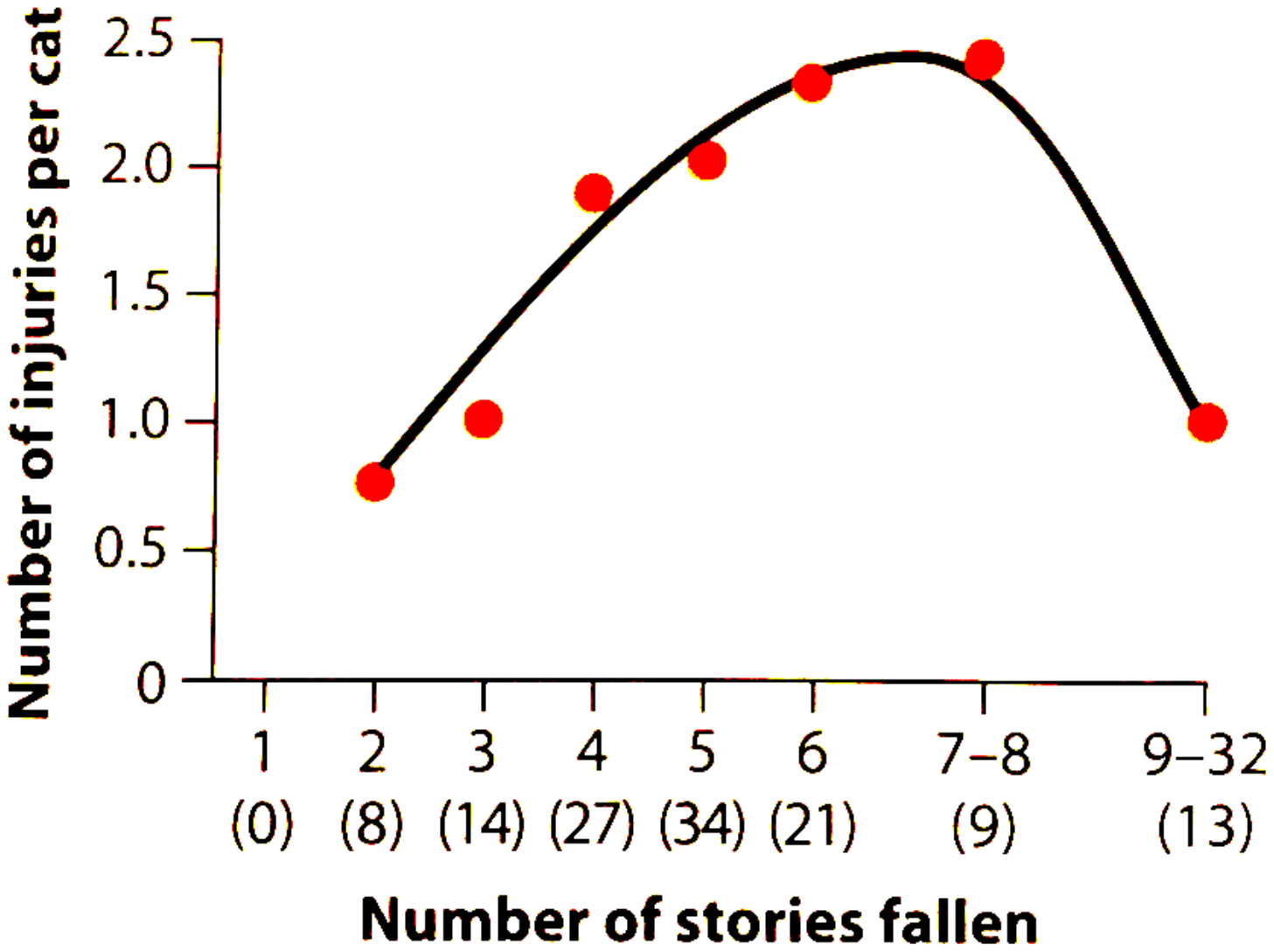
- The group C57BL/6 mice on a high fat diet in the BSBW vivarium
- C57BL/6 mice on a high fat diet
- All *Mus musculus*
- All rodents
- All mammals
- All animals
- All life

Falling Cats

In the period between January 1, 1998 and December 12, 2001 at the Clinic of Surgery, Orthopedics and Ophthalmology of the Veterinary Faculty, 119 cats were treated after a fall or jump from a balcony or window, where the owners saw the fall, or where there was a reasonable suspicion that a fall had occurred. Only those cats that fell from the second or higher stories were included. The owners brought the cats for treatment within varying periods of time after the fall (from 30 min to over a month).

Vnuk, et al. "Feline high-rise syndrome: 119 cases (1998–2001).
Journal of Feline Medicine & Surgery 6.5 (2004): 305-312.

What is the population?



What is your interpretation of the results from this study?

Sampling Considerations

A researcher is interested in understanding how exercise impacts blood pressure. To collect data for their study they go door to door in their neighborhood and ask if any individuals in the house would be interested in participating in the study (a blood pressure reading and a quick questionnaire. To encourage people to participate the researcher is providing $10 Starbucks gift cards.

Random Sampling

1. Every unit in a population should have an equal chance of being sampled.

2. The selection of units must be independent.

3. What are some ways of being non-random?

Hypotheses and P-values

The Null Hypothesis

To analyze your data, you will need a statistical hypothesis to go with your scientific hypothesis

A statistical hypothesis is most easily constructed as a null hypothesis

A null hypothesis posits that the factor of interest has no effect

Frequentist test we will be looking at p-value

Bayesian approaches usually tells us if the posterior estimate of the parameter of interest overlap in our two treatments.

Examples of Null Hypotheses

Fertilizer has no effect on the growth rate of oak trees.

Blocking olfactory cues has no effect on mate choice in swordtail fishes.

Rates of genome evolution are the same in two populations.

Mutations in the 5’ UTR of msl-2 have no effect on translation.

Rejecting the Null

- You apply a statistical test to determine whether your data can reject the null hypothesis

- If you reject the null, then one of the alternative hypotheses must be true – though not necessarily the one you believe or even one you've ever imagined!

- You cannot **prove** a hypothesis, but
  - As frequentist you can find support for an alternative by rejecting the null. The more convincing the null and the more well designed the experiment the more evidence you provide for your alternative.
  - As a Bayesian you can compare support for two competing hypotheses.

What is a p-value?

Is the probability of finding the observed, or more extreme, statistic when the null hypothesis is true (generating the data).

```
> x
     [,1] [,2]
[1,]  140    4
[2,]   80   13
[3,]   76   89
[4,]   20    3
> chisq.test(x)

        Pearson's Chi-squared test

data:  x
X-squared = 117.31, df = 3, p-value < 2.2e-16
```

Number of women on titanic who survived (first column) or died (second column) in first, second, third, or crew classes (rows 1:4 respectively).

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Misconceptions about p-values

The p-value is not the probability that the observed statistic is due to random chance.

A p-value is not the probability that your alternative is false.

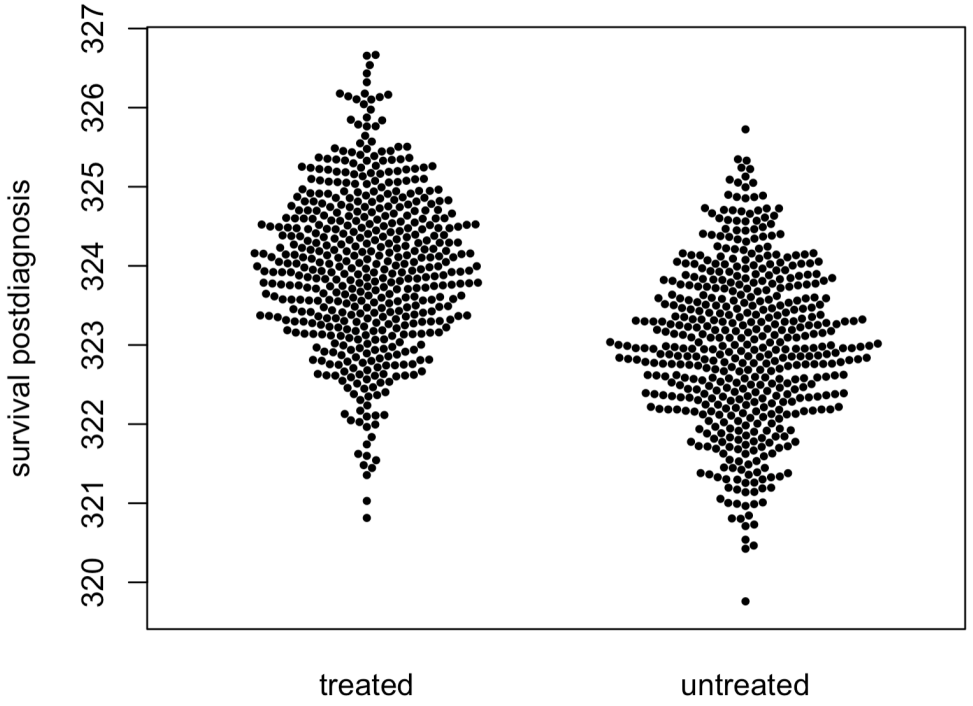
A p-value is not the probability that the null hypothesis is true.

The magnitude of the p-value does not indicate the importance of an effect. Statistical significance does not equate to biological significance.

Studies with p-values on opposite sides of 0.05 are equally “correct”.

Thoughts

A drug company has just developed a new drug Naquadah. This drug is tailored to treat a terminal illness the company reports that their drug increases life span in people who are diagnosed with diseases with a p-value of 1.9×10^{-9}. Unfortunately this drug is very expensive and the prescribed treatment regime would run approximately $40,000 per patient. Should your insurer pay for this drug if you are diagnosed with the diseases.



There are many ways of calculating p-values

Traditional statistical tests

For many questions/experiments there isn't a ready made statistical test.
- Randomization of datasets
- Comparison to simulated datasets

I will ask questions about p-values on tests!

Type I versus Type II Error

Type I error refers to rejecting a true null hypothesis

Type II error refers to failing to reject a false null hypothesis

Power is a description of our probability of rejecting a false null hypothesis

We usually set up statistical tests to avoid Type I errors, at the expense of possibly committing Type II errors

Type I error = FALSE POSITIVE
1 – Type 2 error = POWER

Parameter, estimates, sampling considerations

Parameter: Population-level variables we are trying to estimate

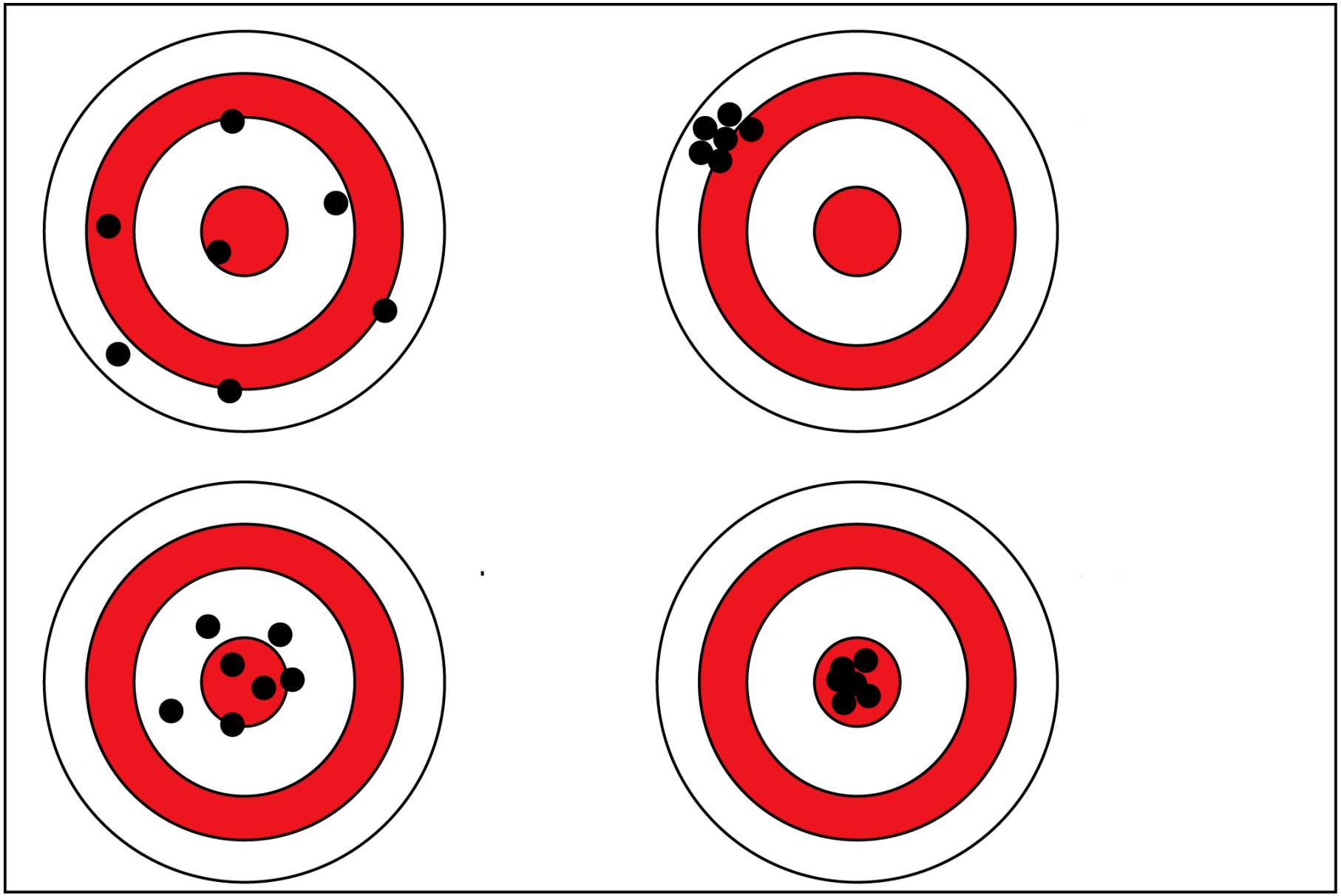
Estimate or Statistic: The value of the parameter inferred from the sample

Bias: If something about the sampling or measuring procedure causes the sample to systematically misrepresent the population.

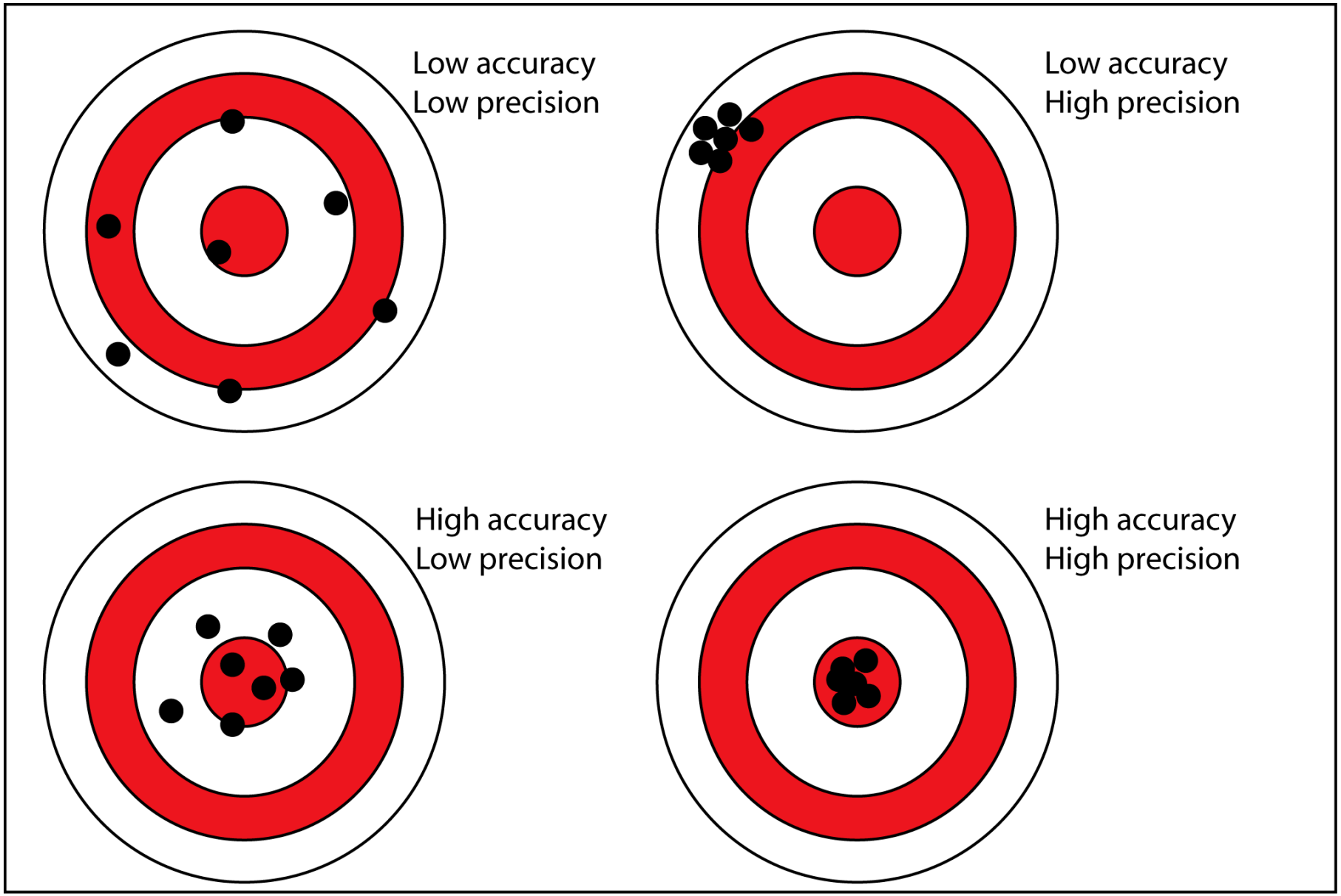
Precision: How tightly grouped are the estimates.

Accuracy: How close estimates are to the true value.

Accuracy vs Precision



Accuracy vs Precision



Experimental vs observational studies

- Does caloric restriction increase lifespan in mice?
- Is global warming caused by human activities?
- Does smoking cause lung cancer in humans?
- Does parasite infection reduce mating success of beetles?
- Does oxytocin affect sexual attraction in humans?
- Do sex chromosomes increase the rate of speciation?
- Do chromosome fusions reduce fitness?

Why should we summarize data?

- Many datasets are simply too big to look at all values and form an impression?

- Our impressions of small datasets are often misled by our tendency to look for patterns.

Typical summary statistics

- **Mean:** Sum of the observations divided by the number of observations
- **Median:** The middle observation in a set of data
- **Variance:** The average squared deviation from the mean
- **Standard Deviation:** The square root of the variance

Symbols for samples and populations

Samples versus Populations

The mean or standard deviation statistic you calculate from your sample is an estimate of the population parameter.

Parameter Symbols:

μ (mu): population mean

σ (sigma): population standard deviation

Statistic Symbols:

$\bar{Y}$ (Y bar): sample mean

s : samples standard deviation

For a sample of a population

The mean is just: $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

The standard deviation is $s = \sqrt{s^2}$

Where s^2 or the variance is: $s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$

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()
```