1. Introduction/Overview

Readings: Benjamin et al (2017)
Goodman (1999)

R: Lab 0 and Lab 1

Overview

A) Introductions
B) Lab/Office Hour Schedule
C) Syllabus
D) What is statistics?
E) Randomness
F) Reproducible Research

My Own Intro

Assistant Professor of Biostatistics and Informatics in the Colorado School of Public Health

My own research is on Bayesian adaptive clinical trial designs and Bayesian methods to facilitate the sharing of information across multiple sources

I have a big focus on collaborative research through the Center for Innovative Design and Analysis (CIDA) in pediatric obesity, gastroenterology, otolaryngology, and anesthesiology

I have one rescue named Baisy (whom my family named because it was the closest thing they could remember from my dissertation)

Introductions

We will go around the room and do quick introductions so I can start to put names to faces!

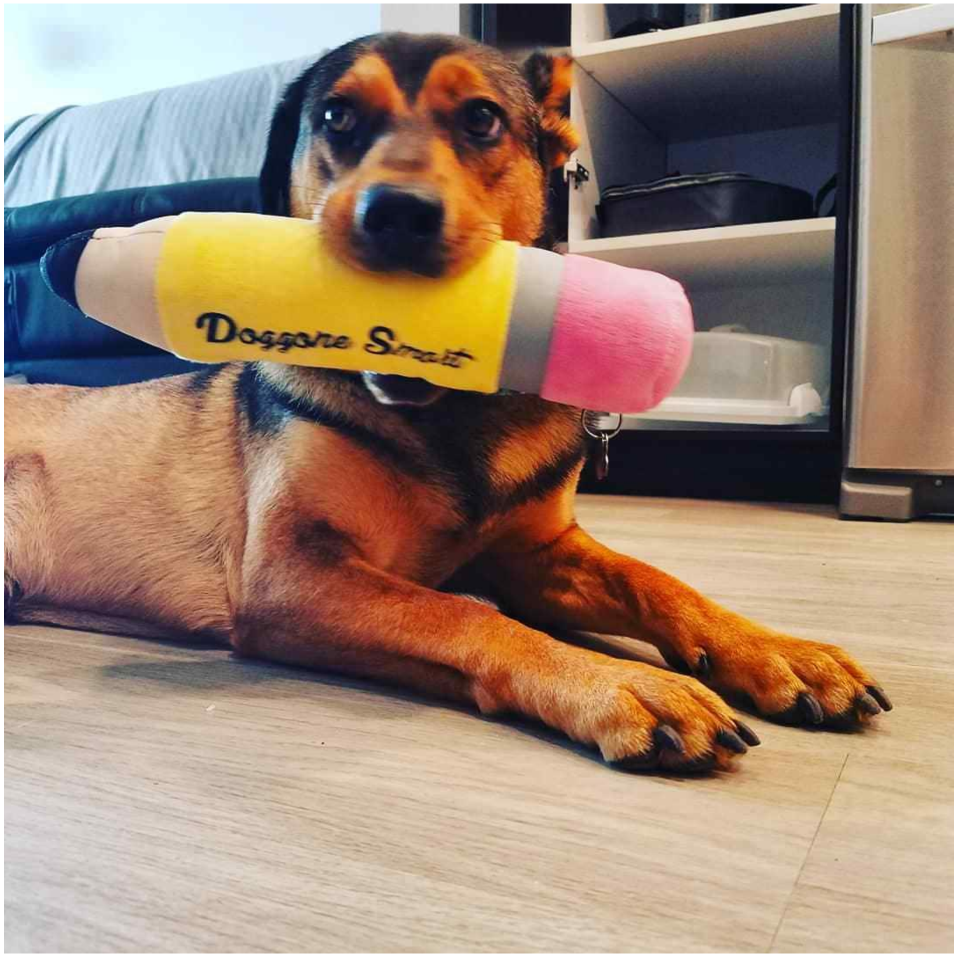
Include the following:

1. Name
    a. Alex Kaizer

2. Program
    a. PhD in Biostatistics

3. 1 fun thing you did this summer
    a. I finally got a new rear bike tire…after 6 flats this summer…

Lab/Office Hours Schedule and Syllabus



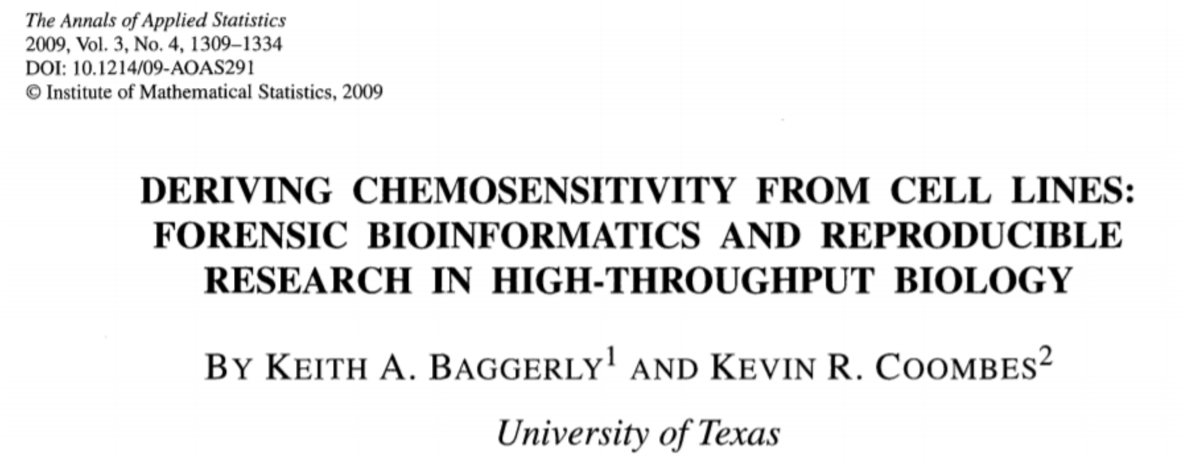
What is statistics/biostatistics?

- Statistics as a discipline arose from the need to use data to answer scientific questions in the face of uncertainty

- Statistical concepts are at the heart of scientific inquiry in the health sciences

- Your mastery of fundamental concepts will facilitate:
    - a better understanding of published research
    - a better understanding of how to structure effective scientific research
    - interpretation and presentation of results
    - collaboration with other biostatisticians and scientific investigators

Compelling example that underscores the need for careful design, data handling, and analysis:

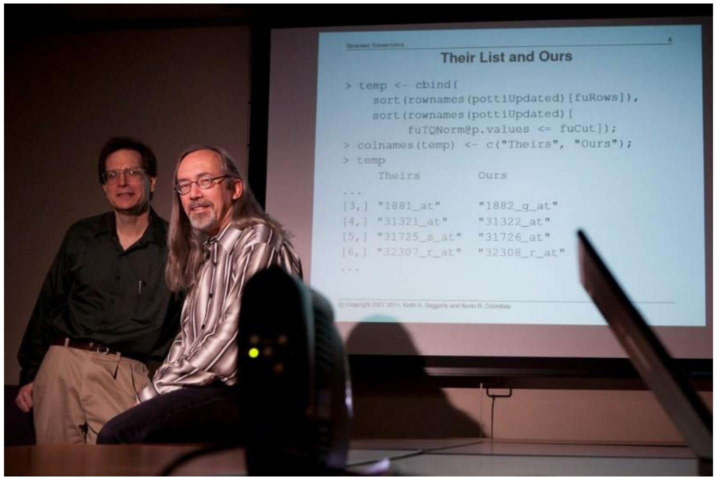


The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1309–1334
DOI: 10.1214/09-AOAS291
© Institute of Mathematical Statistics, 2009

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY[1] AND KEVIN R. COOMBES[2]

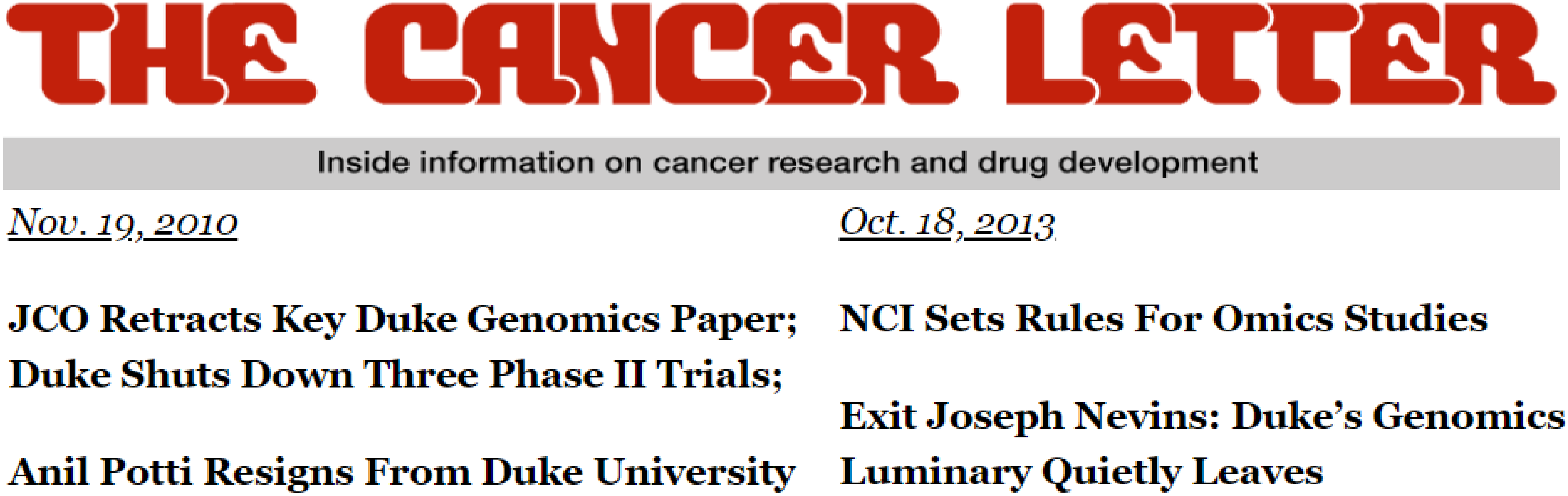
University of Texas



Their List and Ours

```
> temp <- cbind(
    sort(rownames(pottiUpdated)[fuRows]),
    sort(rownames(pottiUpdated)[
         fuTQNorm@p.values <= fuCut]);
> colnames(temp) <- c("Theirs", "Ours");
> temp
      Theirs         Ours
...
[3,] "1881_at"      "1882_g_at"
[4,] "31321_at"     "31322_at"
[5,] "31725_s_at"   "31726_at"
[6,] "32307_r_at"   "32308_r_at"
...
```

Source: New York Times



THE CANCER LETTER

Inside information on cancer research and drug development

Nov. 19, 2010

JCO Retracts Key Duke Genomics Paper;
Duke Shuts Down Three Phase II Trials;

Anil Potti Resigns From Duke University

Oct. 18, 2013

NCI Sets Rules For Omics Studies

Exit Joseph Nevins: Duke's Genomics Luminary Quietly Leaves

Algorithms vs. Inference

Types of algorithms

Dip the lasagna sheets in boiling water + salt + olive oil
prevent pasta from sticking
Drain
put the lasagna sheets in cold water
100% Virgin (olive oil)
Olive Oil
3rd stage • Boil the lasagna sheets
breadcrumbs
put all the ingredients together
4th stage • Assemble & Bake
Layers
1 - tomato sauce
2 - lasagna sheets
3 - tomato sauce mushrooms green peas radicchio & béchamel sauce
Repeat 1 - 2 - 3
On top: tomato sauce and breadcrumbs
Finally put your Vegan Pasticcio in the oven - preheated at 180C - for 45 min.
ISIO'13

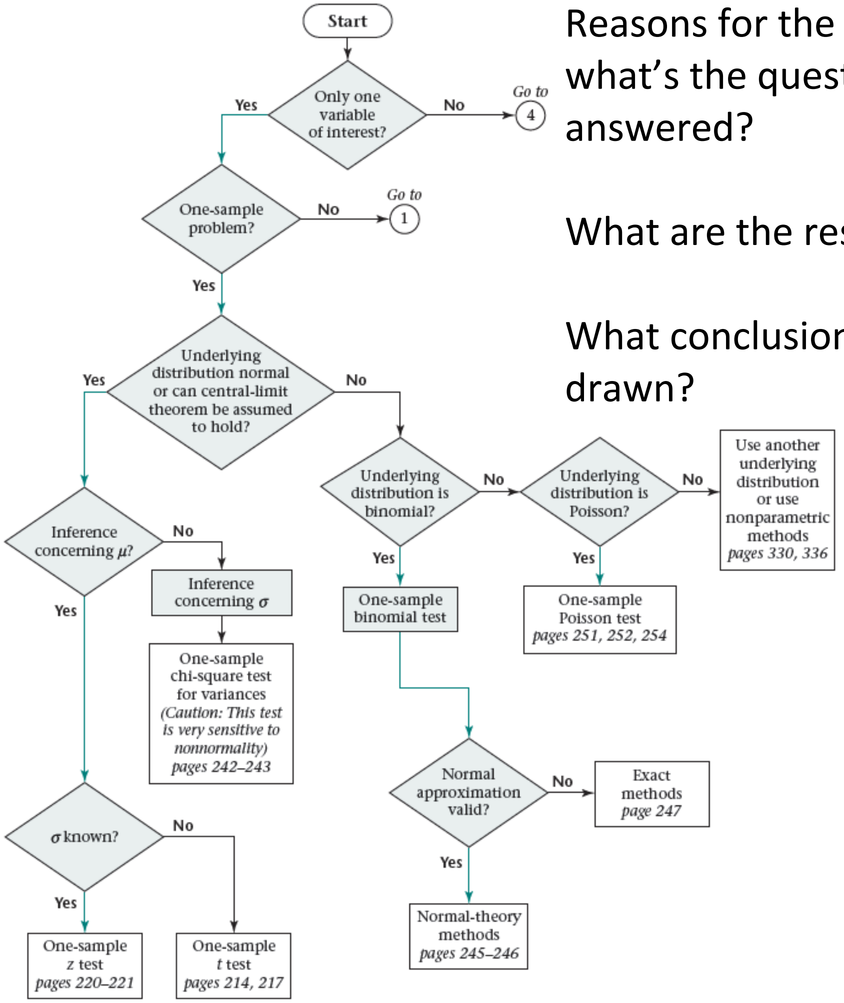
Inference

Inference

Reasons for the algorithm – what's the question being answered?

What are the results?

What conclusions can be drawn?


Start

Only one variable of interest?

Yes

No

Go to 4

One-sample problem?

No

Go to 1

Yes

Underlying distribution normal or can central-limit theorem be assumed to hold?

Yes

No

Inference concerning μ?

No

Inference concerning σ

One-sample chi-square test for variances (*Caution: This test is very sensitive to nonnormality*) *pages 242–243*

Yes

σ known?

Yes

No

One-sample *z* test *pages 220–221*

One-sample *t* test *pages 214, 217*

Underlying distribution is binomial?

No

Yes

One-sample binomial test

Normal approximation valid?

No

Exact methods *page 247*

Yes

Normal-theory methods *pages 245–246*

Underlying distribution is Poisson?

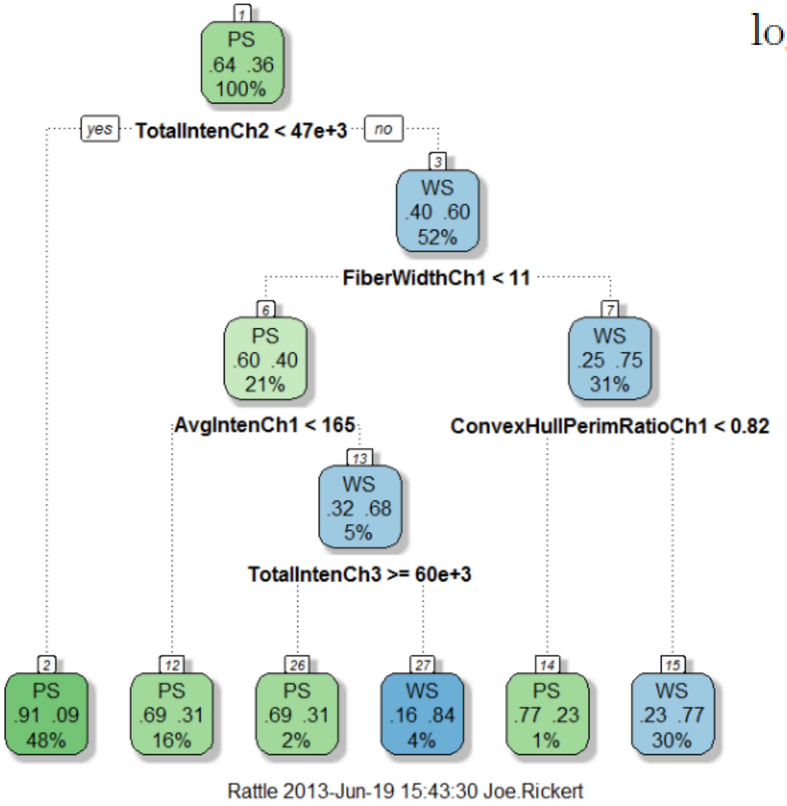
No

Use another underlying distribution or use nonparametric methods *pages 330, 336*

Yes

One-sample Poisson test *pages 251, 252, 254*

Algorithms vs. Models



$$
\begin{aligned}
\text{logit}(\mathbb{E}[Y_i|x_{1,i}, \ldots, x_{m,i}]) &= \text{logit}(p_i) \\
&= \log\left(\frac{p_i}{1-p_i}\right) \\
&= \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i}
\end{aligned}
$$

$$
\begin{aligned}
\text{logit}(\mathbb{E}[Y_i|\mathbf{X}_i]) &= \text{logit}(p_i) \\
&= \log\left(\frac{p_i}{1-p_i}\right) \\
&= \beta\mathbf{X}_i
\end{aligned}
$$

Randomness

Randomness is fundamental to statistical inference

Control of randomness is key to experimentation and the scientific method

Random Number Generation

True Random Number Generator (TRNG) vs. Pseudo Random Number Generator (PRNG)

Characteristic	**PRNG**	**TRNG**
Efficiency	Excellent	Poor
Determinism	Deterministic	Nondeterministic
Periodicity	Periodic	Aperiodic

Application	**Most Suitable Generator**
Lotteries and Draws	TRNG
Games and Gambling	TRNG
Random Sampling (e.g., drug screening)	TRNG
Simulation and Modelling	PRNG
Security (e.g., generation of data encryption keys)	TRNG
The Arts	Varies

Source: www.random.org/randomness

Applications of random number generation

A. Research Design

Randomization/random allocation, random sampling

1. **Random Digits:** (Combinations of) 0, 1, …, 9 occur with the same relative frequency (uniformly distributed).
   a. Digits occur independently of occurrence of other digits
   b. Use these sequences for random selection or allocation
   c. For example, a random number table

2. **Random Selection (Sampling):** Selecting random portion of large population (e.g., select 10 units randomly from 1000)
   a. Tools in R include the “runif” and “sample” functions
   b. Simple random sampling, cluster sampling, stratified sampling

3. **Random Allocation:** Assigning treatments randomly to individual units or groups of units
   a. Tools in R include the “blockTools” package

Applications of random number generation

B. Random Sampling from Theoretical Distributions

Also known as *simulation*

Overview

- Simulation is a fundamental and powerful tool in statistical practice
- Simulation for understanding
- Simulation for experimentation
- https://www4.stat.ncsu.edu/~davidian/st810a/simulation_handout.pdf

IRReproducibility

Reliance on p-values –

- long-term trends leading up to the data + evidence provided by data; conflation of these has resulted in current approach which generally satisfies neither (Goodman, 1999)
- < 0.05 actually provides weak evidence against H_0 (Benjamin et al. 2017)

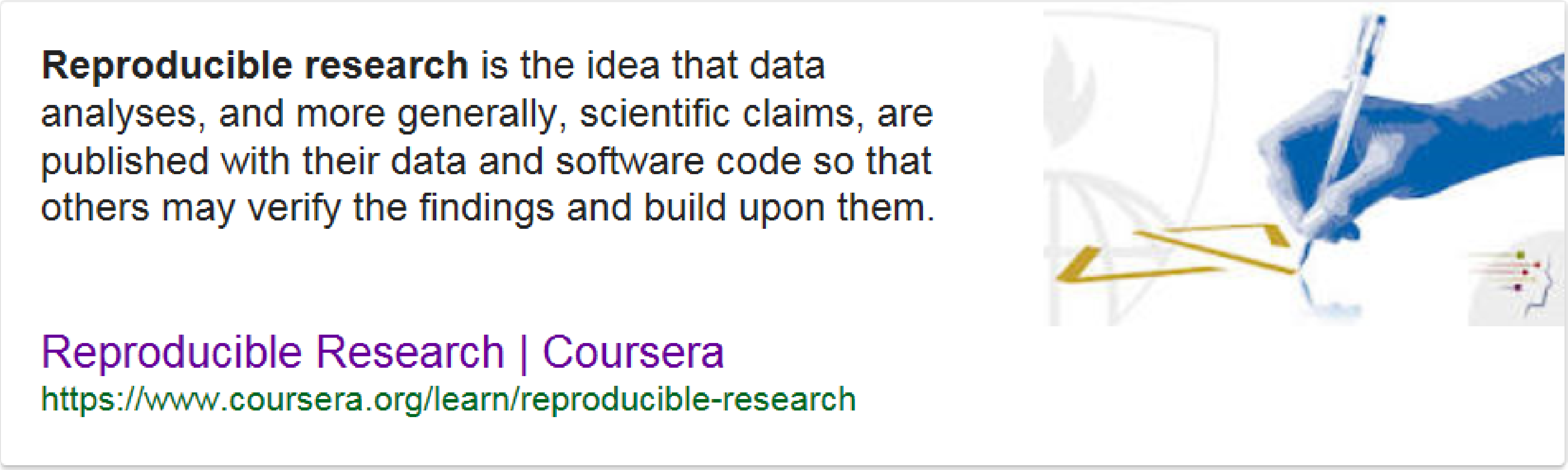
Multiplicity – leads to *selection* of results and this has an *effect* on observed ability to replicate results (see 2005 *JAMA* paper by Ioannidis on course website)

- Multiple variables, endpoints, time points, subgroups, comparisons
- Multiple hypothesis testing, multiple looks at the data
- Multiple models and adjustments
- Fishing expeditions, mountains of output without *a priori* thought and justification - *…what exactly <u>was</u> my (their) research hypothesis or question?…p-hacking …*

Publication bias – file drawer problem, publish or perish, sensationalism

Cognitive biases – preconceived notions about what effects are real, what effects could be real, what effects are likely not to be real

Reproducible Research



Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them.

Reproducible Research | Coursera

https://www.coursera.org/learn/reproducible-research

Open Science Paradigm

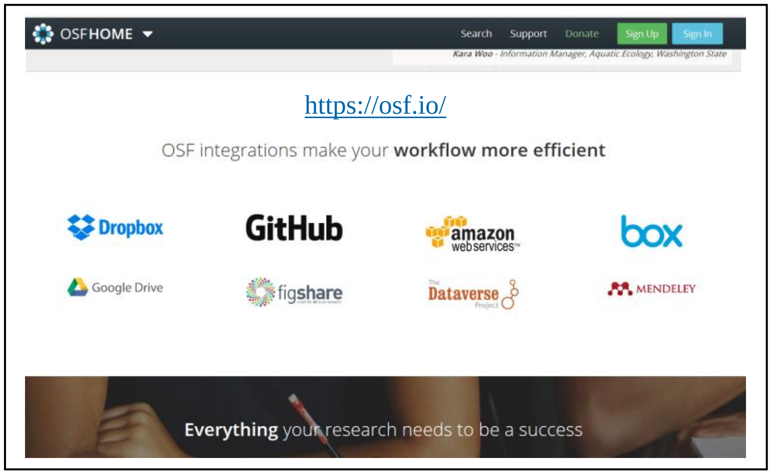
Open Access - Publications

Public access – free to anyone with an internet connection

Free: Use, reuse, remix

Pre-study: Study registration

Post-study: Preprints



Kara Woo - Information Manager, Aquatic Ecology, Washington State

https://osf.io/

OSF integrations make your **workflow more efficient**

Everything your research needs to be a success

Open Data

Known Provenance

Confidentiality assured

Portability (interoperability) built in

Excel-free, reproducible data manipulation/management best practices applied

Open Code

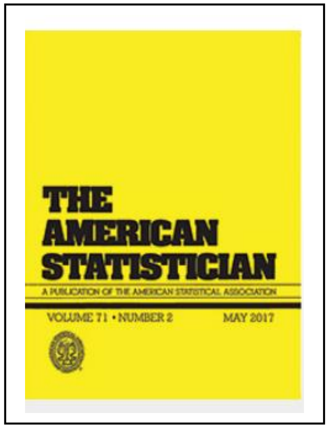
Version control used to track changes

Collaborative model – team science

Crowd sourcing – best solutions openly available

Miscellaneous

Throughout semester we'll:


THE AMERICAN STATISTICIAN
A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION
VOLUME 71 • NUMBER 2
MAY 2017

Read

The statistical methods literature

Practice Reproducible Research Principles

"Lite" version …move towards compliance

Appreciate

("Peak") science writing – WIRED, Quanta, your favorites

Be on the lookout for cool, innovative graphics

ggjoy – R package

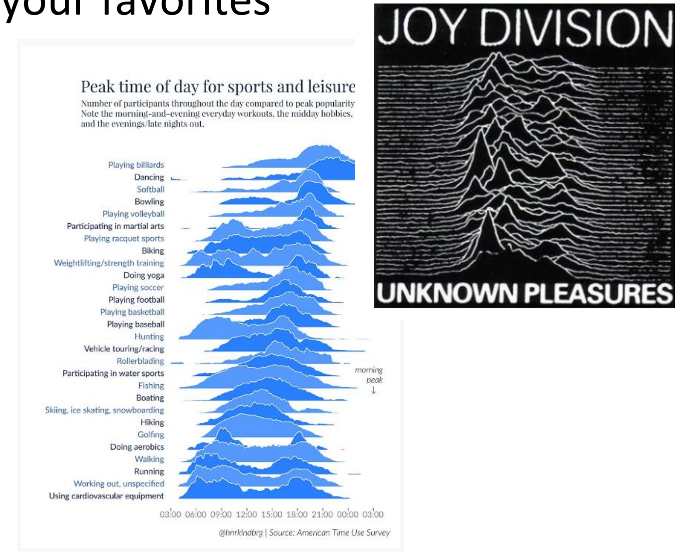
https://cran.r-project.org/web/packages/ggjoy/vignettes/introduction.html

R Gallery:

https://cran.r-project.org/web/packages/ggjoy/vignettes/gallery.html

Share

Resources …using Canvas


Peak time of day for sports and leisure
Number of participants throughout the day compared to peak popularity
Note the morning-and-evening everyday workouts, the midday hobbies, and the evenings/late nights out.
Playing billiards
Dancing
Softball
Bowling
Playing volleyball
Participating in martial arts
Playing racquet sports
Biking
Weightlifting/strength training
Doing yoga
Playing soccer
Playing football
Playing basketball
Playing baseball
Hunting
Vehicle touring/racing
Rollerblading
Participating in water sports
Fishing
Boating
Skiing, ice skating, snowboarding
Hiking
Golfing
Doing aerobics
Walking
Running
Working out, unspecified
Using cardiovascular equipment
morning peak
03:00 06:00 09:00 12:00 15:00 18:00 21:00 00:00 03:00
@hnrklndbrg | Source: American Time Use Survey
JOY DIVISION
UNKNOWN PLEASURES