Quick overview of basic concepts in Statistics

(and a few bad jokes)

Motivation

Statistical concepts can be subtle and difficult to explain

I thought it would be useful to review some central ideas in Statistics

Overview

Descriptive statistics

- Distributions, sampling
- Descriptive statistics and plots
- Sampling distributions

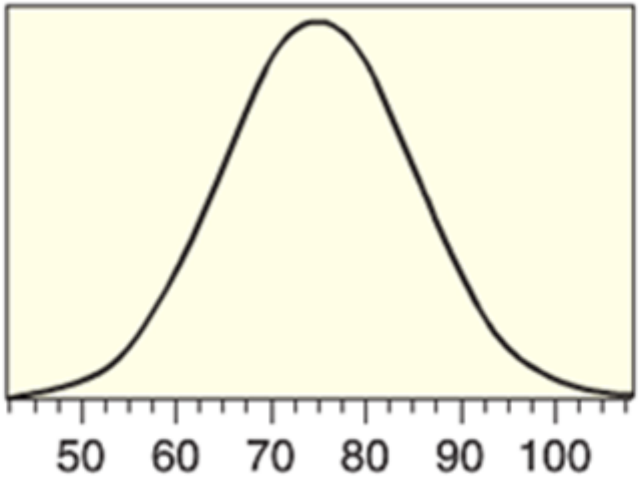
Inferential statistics

- Point and interval estimates
- Confidence intervals and the bootstrap
- Hypothesis tests

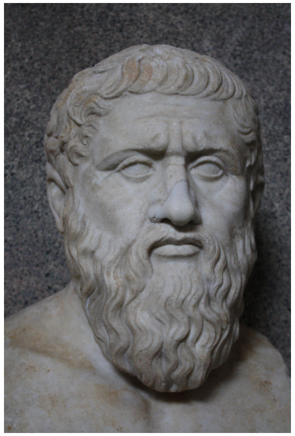
If there is interest: neural data analysis

Where do data come from?

Distributions!



Data ☹



Truth!

How do we get data?

Simple random sample: each member in the population is equally likely to be in the sample

- Random selection

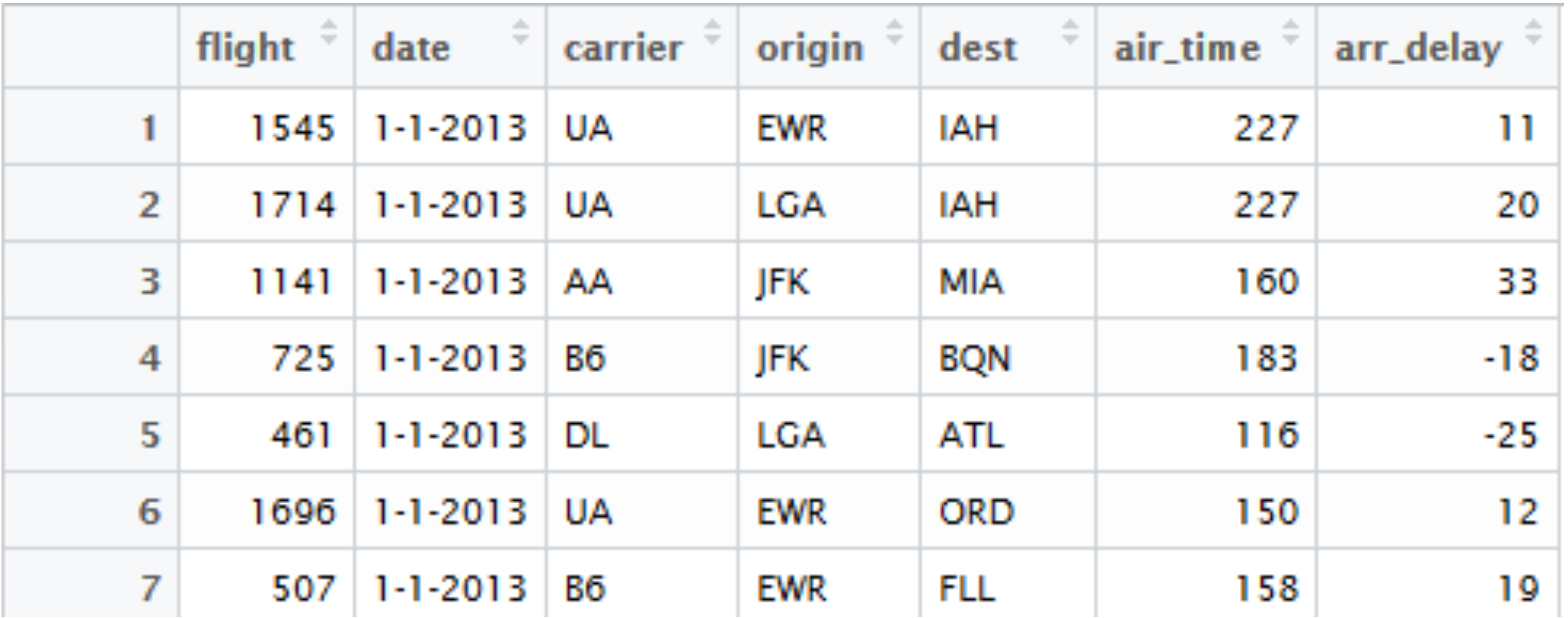
Soup analogy!

Q: Why is this good?

An Example Dataset (flight delays)

Variables

Cases



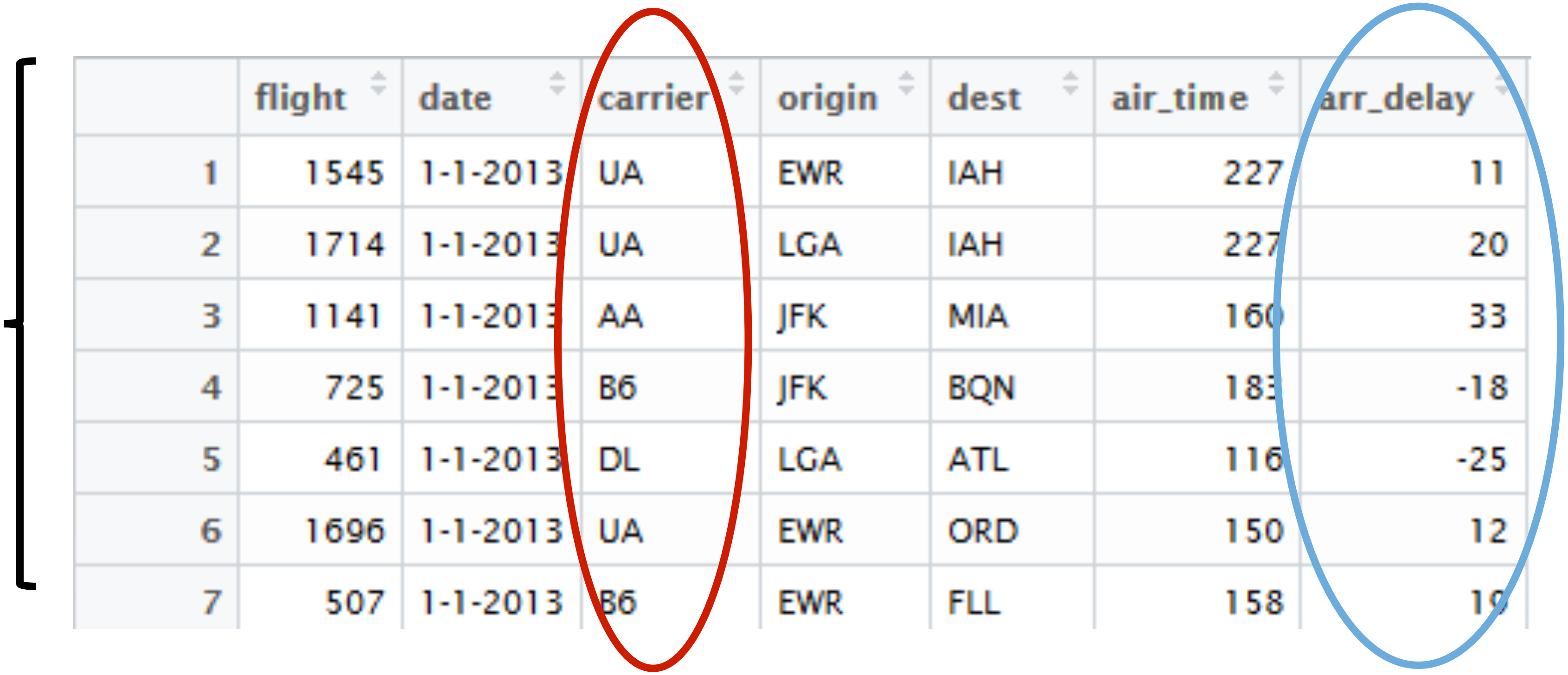
	flight	date	carrier	origin	dest	air_time	arr_delay
1	1545	1-1-2013	UA	EWR	IAH	227	11
2	1714	1-1-2013	UA	LGA	IAH	227	20
3	1141	1-1-2013	AA	JFK	MIA	160	33
4	725	1-1-2013	B6	JFK	BQN	183	-18
5	461	1-1-2013	DL	LGA	ATL	116	-25
6	1696	1-1-2013	UA	EWR	ORD	150	12
7	507	1-1-2013	B6	EWR	FLL	158	19

An Example Dataset (flight delays)

Categorical Variable

Quantitative Variable

Cases



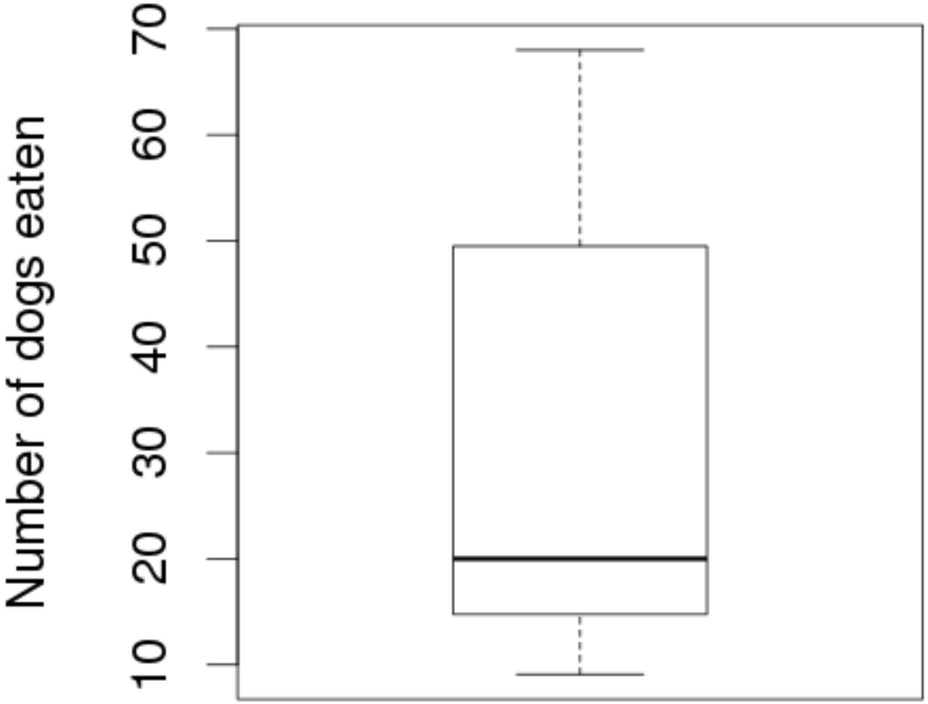
	flight	date	carrier	origin	dest	air_time	arr_delay
1	1545	1-1-2013	UA	EWR	IAH	227	11
2	1714	1-1-2013	UA	LGA	IAH	227	20
3	1141	1-1-2013	AA	JFK	MIA	160	33
4	725	1-1-2013	B6	JFK	BQN	183	-18
5	461	1-1-2013	DL	LGA	ATL	116	-25
6	1696	1-1-2013	UA	EWR	ORD	150	12
7	507	1-1-2013	B6	EWR	FLL	158	19

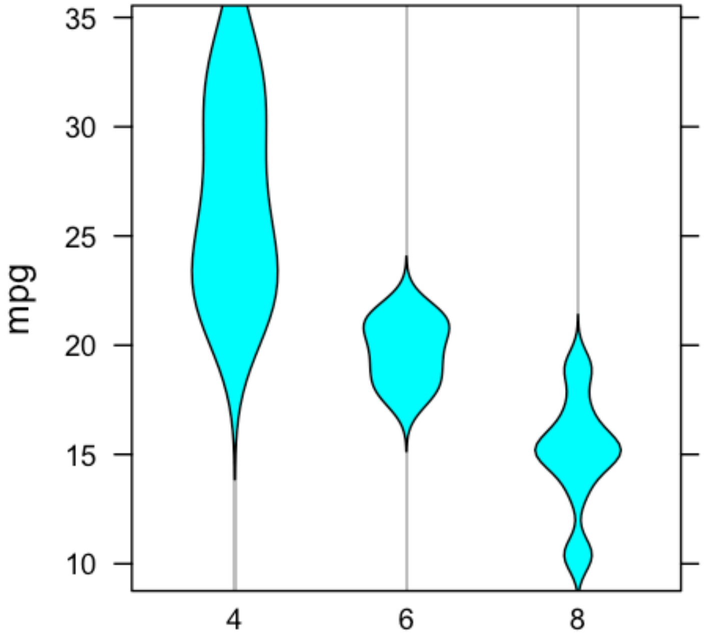
What is a good first step when analyzing data?

What is a useful plot
for categorical data?

What is a useful plot
for quantitative data?

Box and violin plots

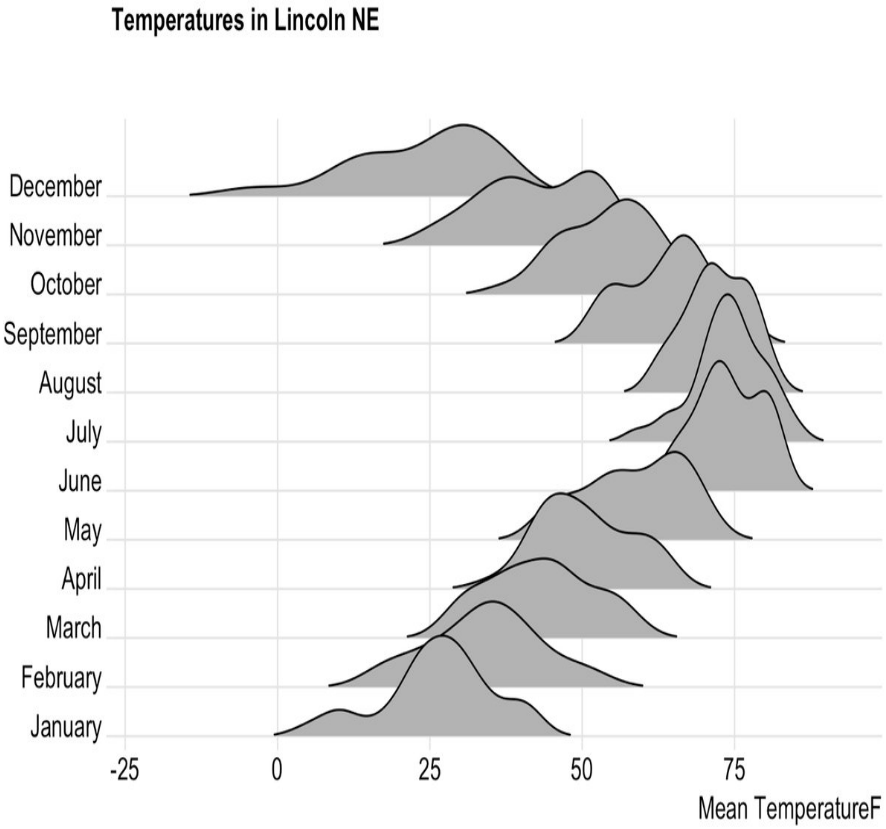




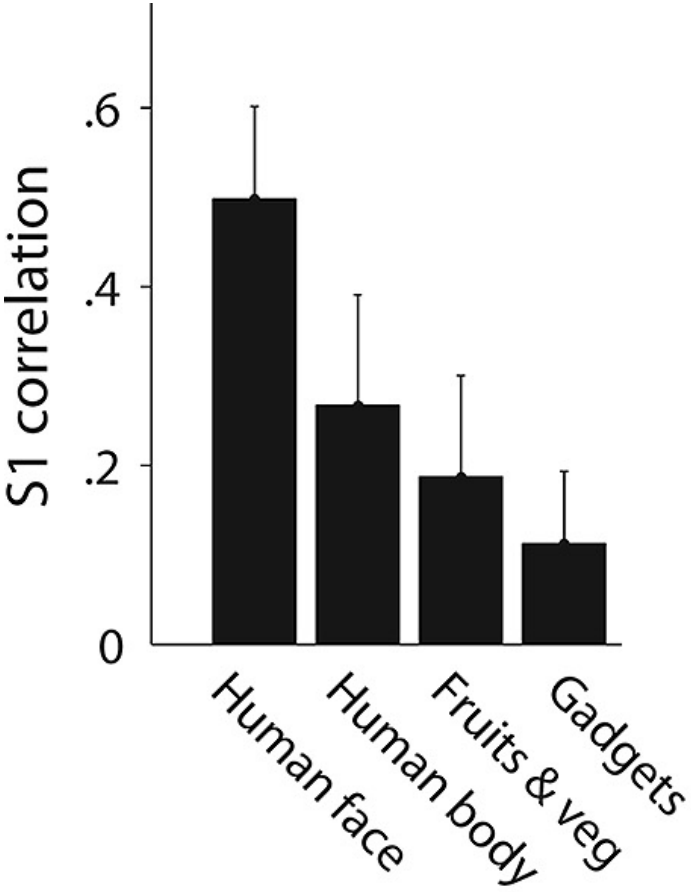
Joy plots

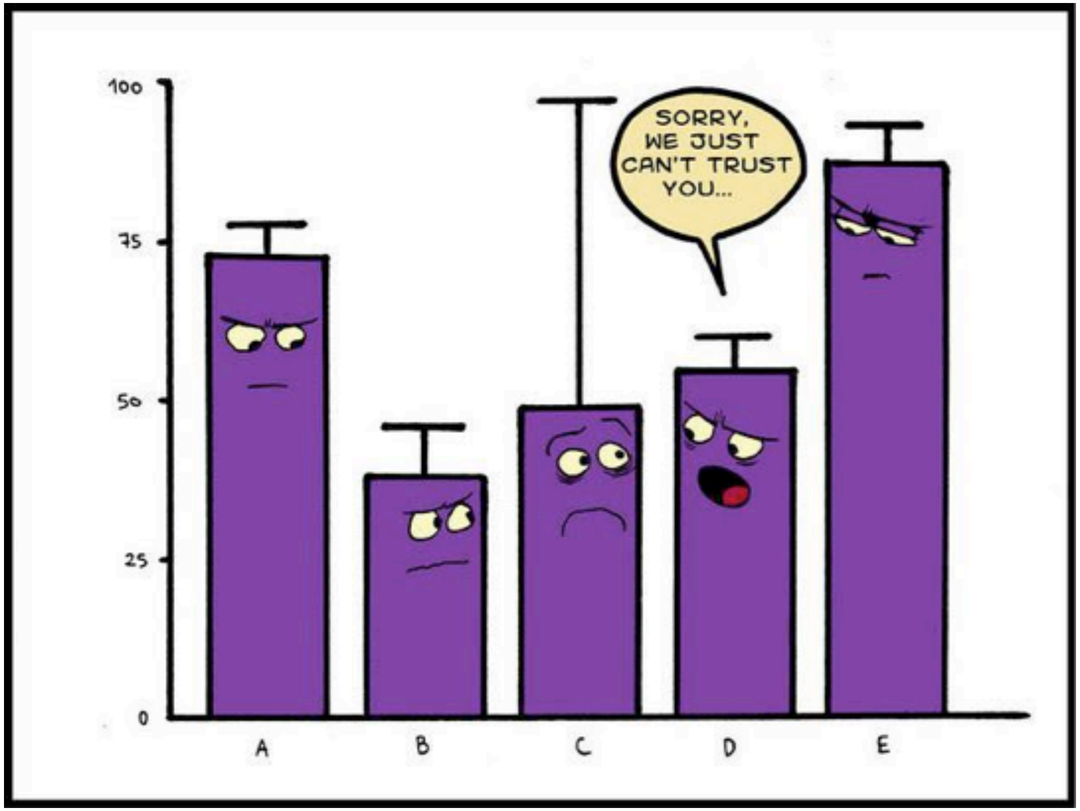
Neat fact: Some statisticians find violin plots ugly

- Why do they find them ugly?



Dynamite plots





Neat fact: Many statisticians hate dynamite plots

- Why do they hate them?

What is a statistic?

A: A single death is a tragedy; a million deaths is a statistic.

- Joseph Stallin

Descriptive statistics describe the data you have collected

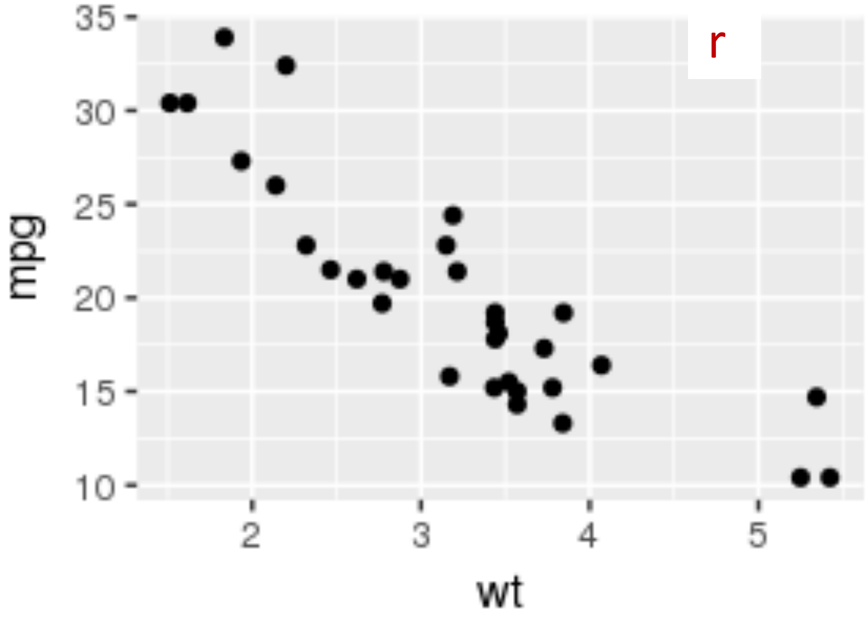
- Usually denoted with roman characters
  - A single categorical variable: proportion
  - A single quantitative variable: the mean
  - A pair of quantitative variables: the correlation

Example statistic: correlation coefficient

The correlation coefficient *r* is a statistic that measures the linear association between two variables X, and Y

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

r is an estimate of ρ



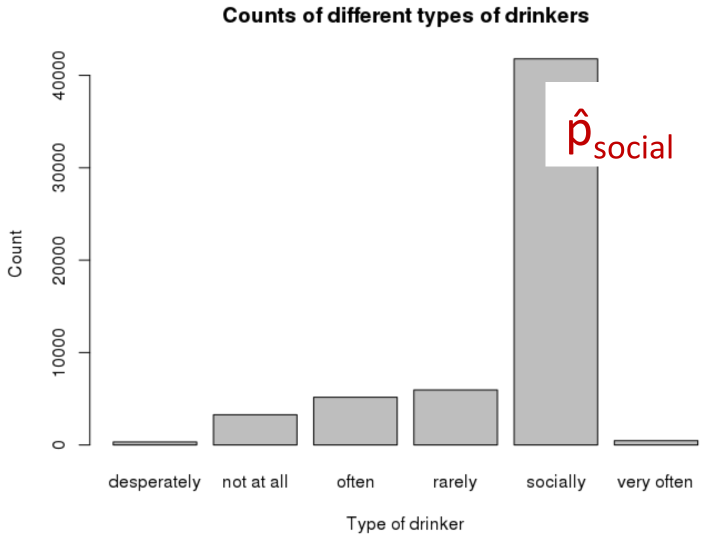
Side note: correlation ≠ causation

The correlation coefficient *r* is a statistic that measures the linear association between two variables X, and Y

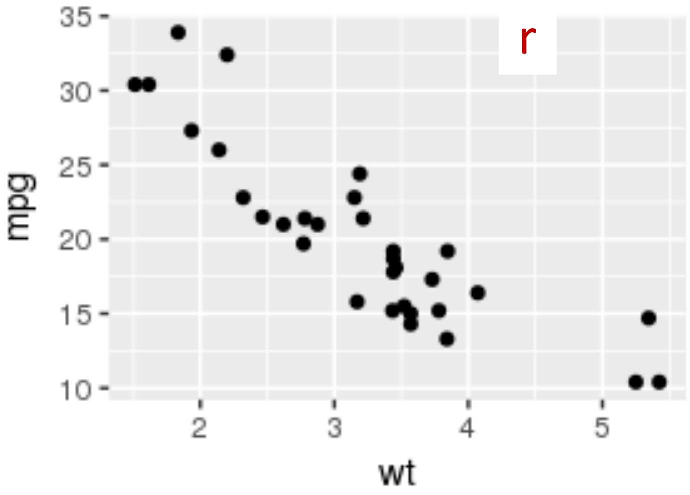
A variety of statistics than can be used to describe data

Descriptive ***statistics*** are usually denoted with roman characters

- A single quantitative variable: the mean
- A single categorical variable: proportion
- A pair of quantitative variables: the correlation



$\hat{p}_{social}$

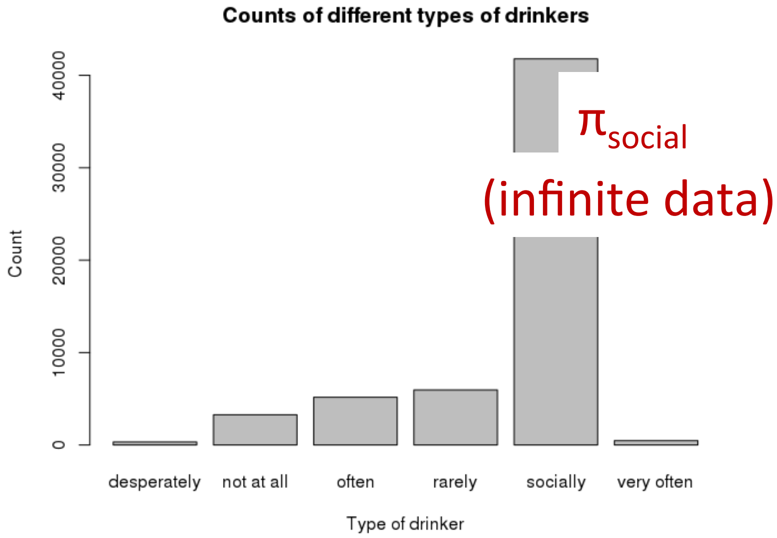


r

Statistics have corresponding parameters

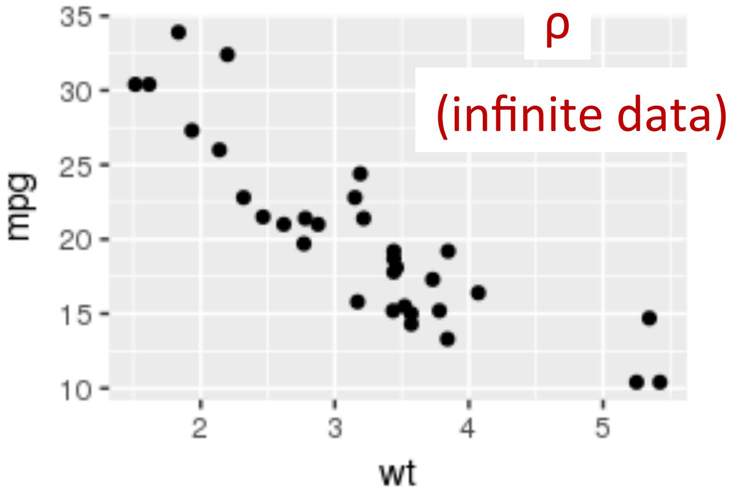
Population/process ***parameters*** are usually denoted with greek characters

- A single quantitative variable: the mean
- A single categorical variable: proportion
- A pair of quantitative variables: the correlation



π_{social}

(infinite data)

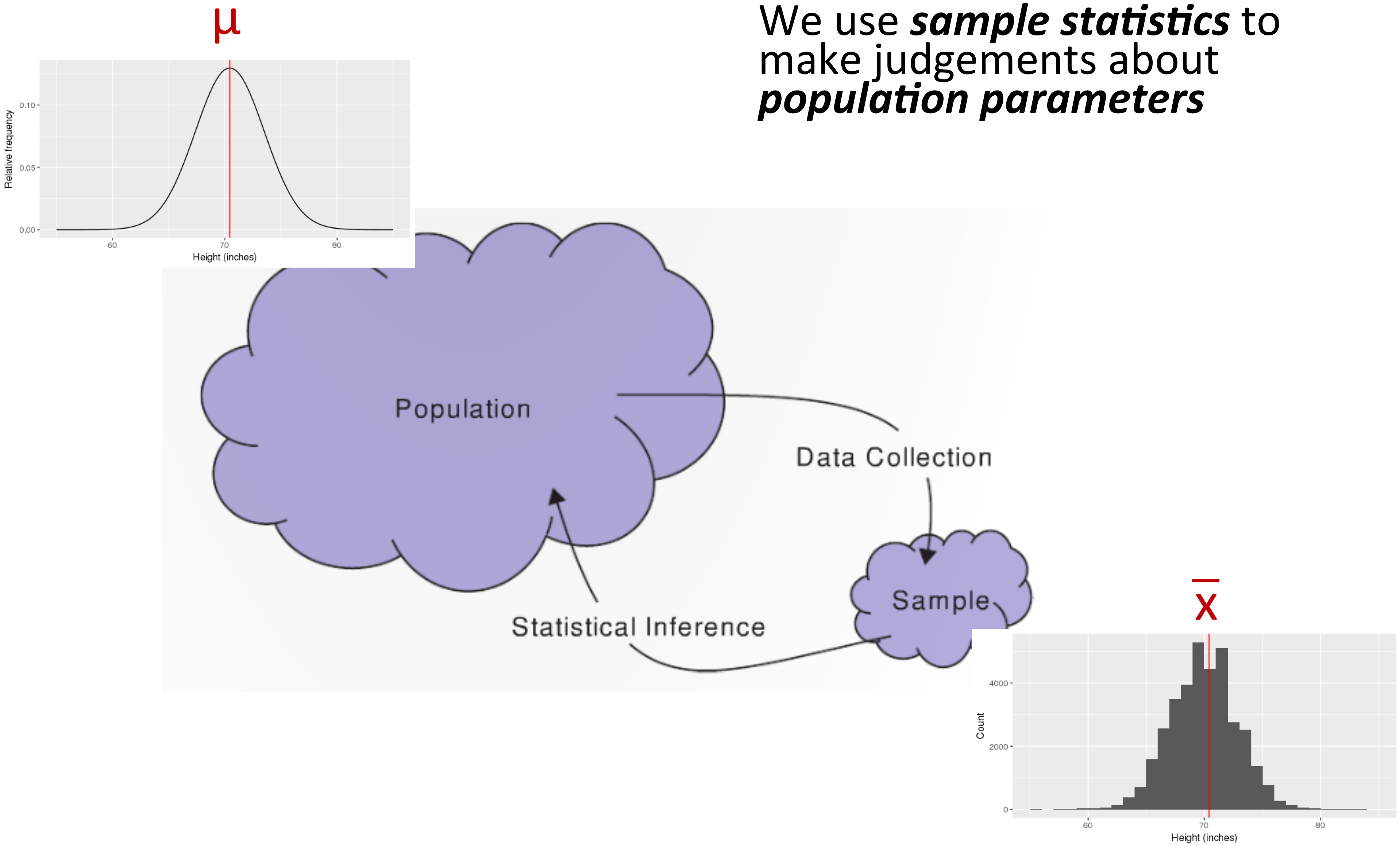


ρ

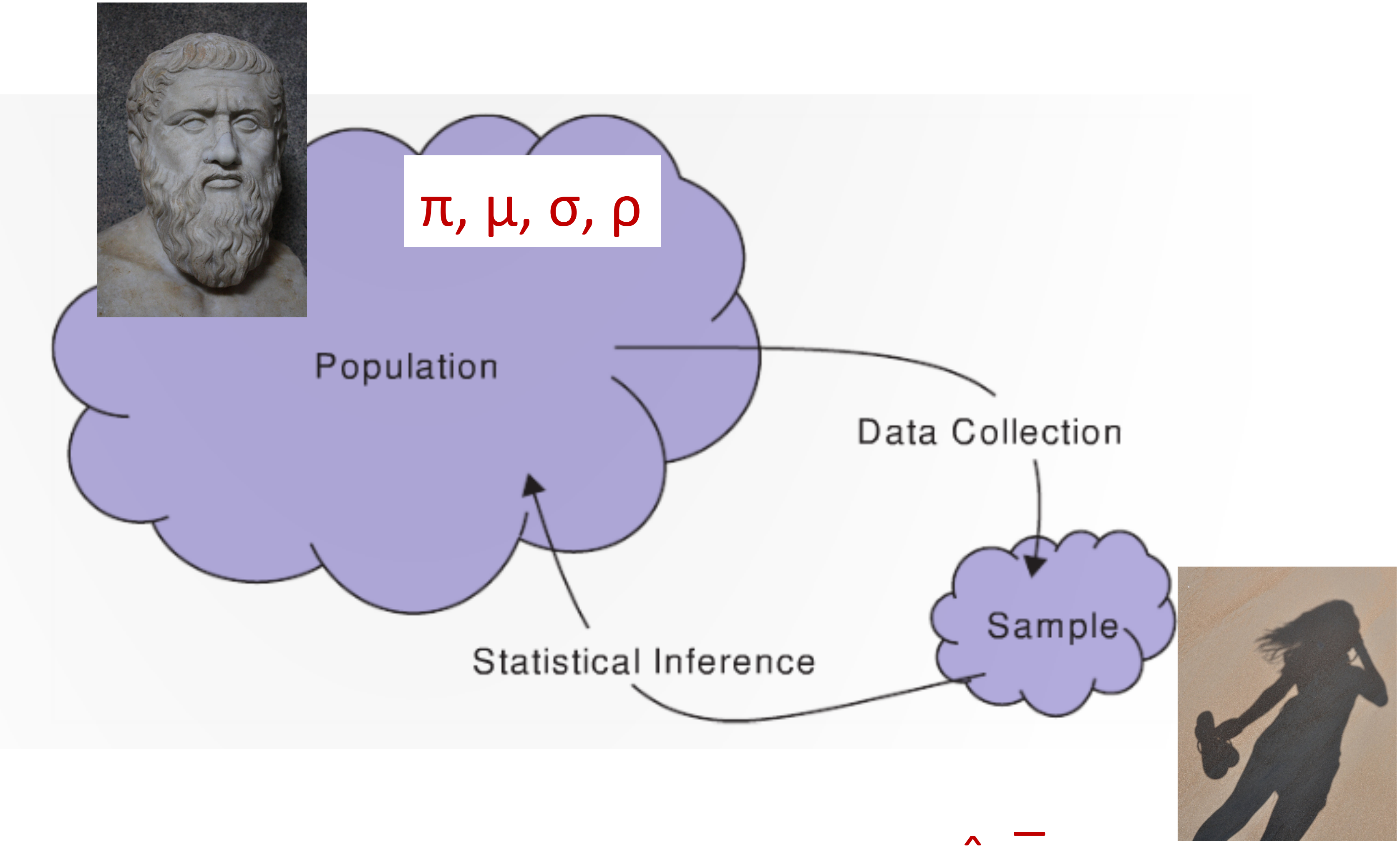
(infinite data)

Statistical inference

We use ***sample statistics*** to make judgements about ***population parameters***



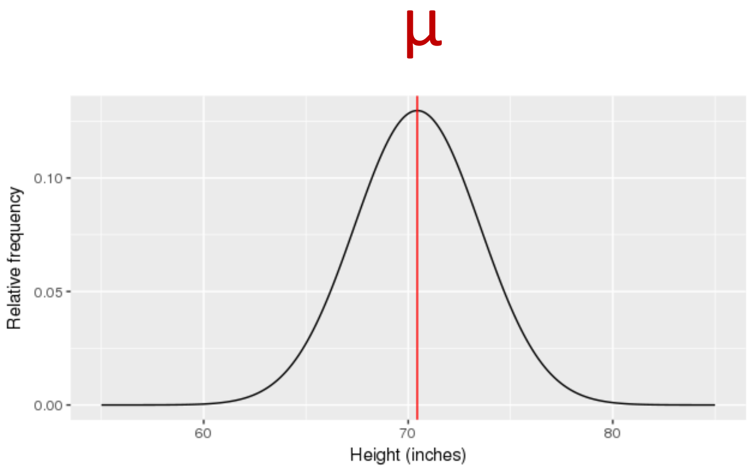
Statistical inference

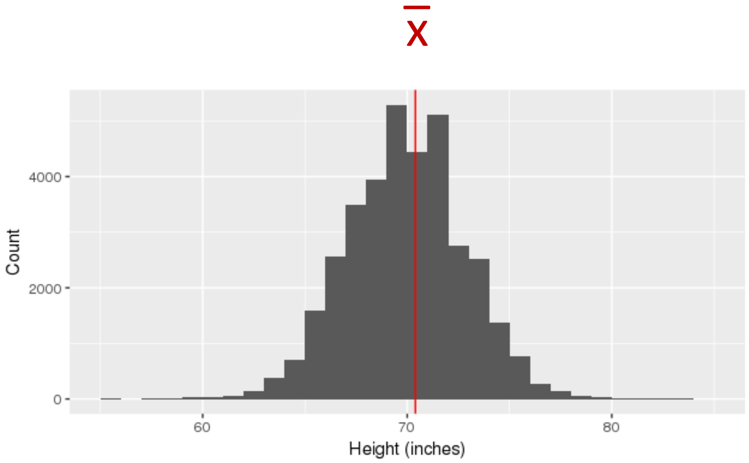


$\hat{p}$, $\bar{x}$, s, r

Estimation

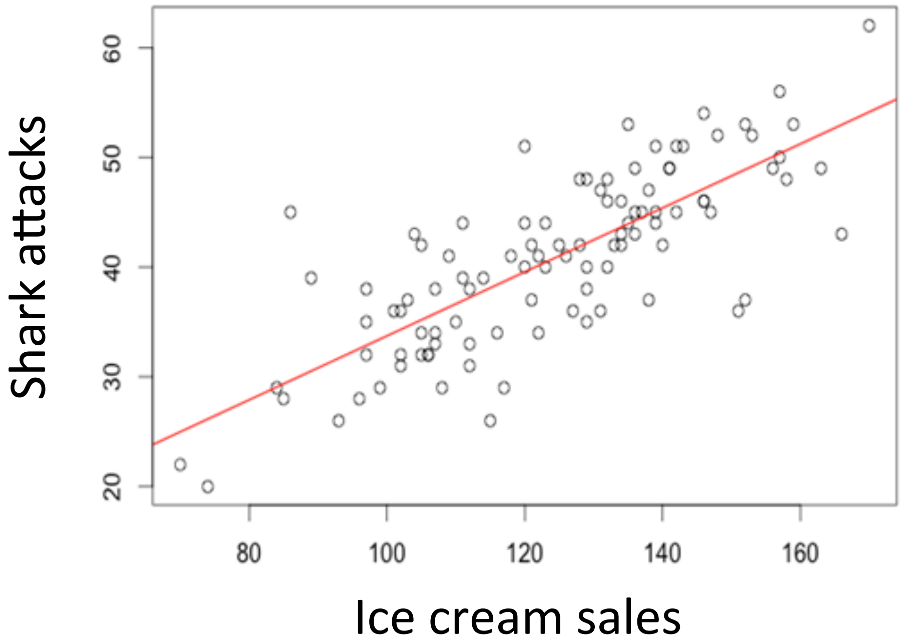
Point estimation: $\bar{x}$ is a point estimate for μ





Regression

In linear regression, we try to fit a line to our data



Truth: $y \;=\; \beta_0 \;+\; \beta_1 \cdot x$

Fit: $\hat{y} \;=\; b_0 \;+\; b_1 \cdot x$

The b_i's are usually chosen to minimize the MSE:

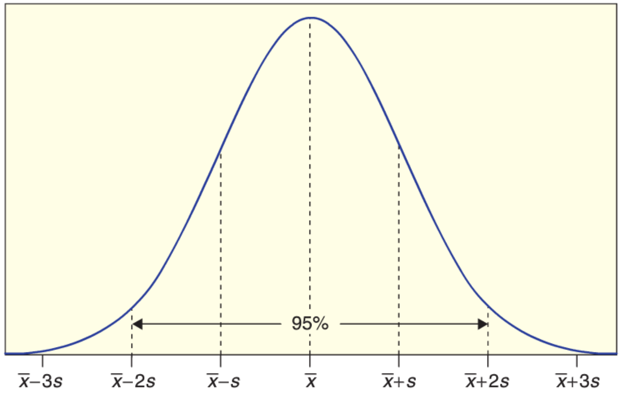
$$MSE \;=\; \frac{1}{n}\sum_i^n (\hat{y}_i - y_i)^2$$

Regression caution # 1

Avoid trying to apply the regression line to predict values far from those that were used to create the line

Statistics have distributions too!

A **sampling distribution** is the distribution of sample statistics computed for different samples of the same size from the same population $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_k$



Sampling distribution are often approximately Normal

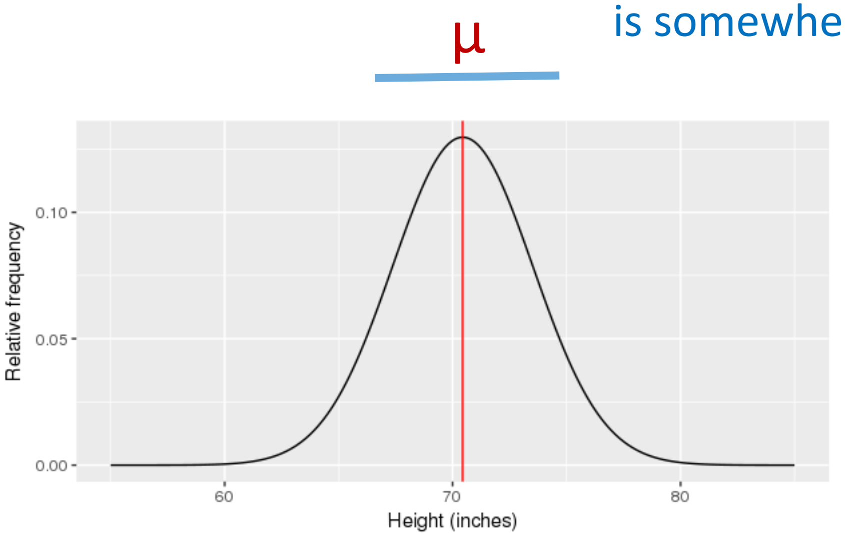
- Due to the central limit theorem: sums of many random variates lead to a Normal distribution

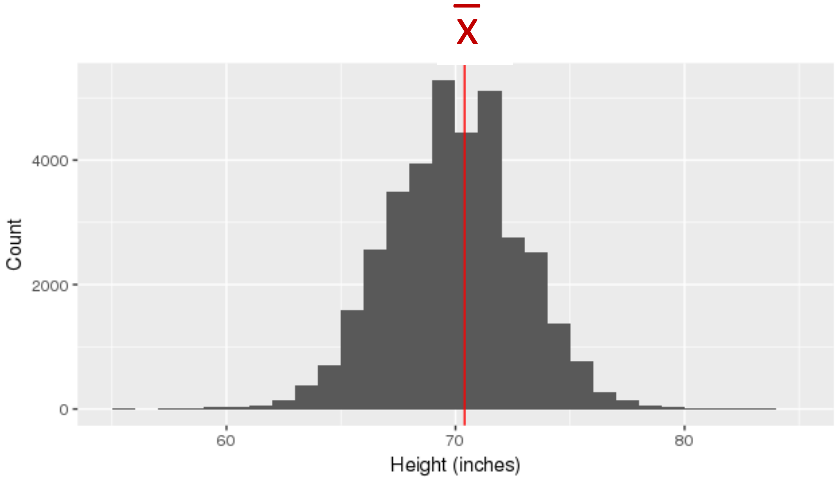
Estimation

Point estimation: $\bar{x}$ is a point estimate for μ

Interval estimate: point estimate + margin of error

The population mean is somewhere in here





Confidence Intervals

A **confidence interval** *for a parameter* is an interval computed by ***a method that will capture the parameter a specified proportion of times***

- (if the sampling process were repeated many times)

The success rate (proportion of all samples whose intervals contain the parameter) is known as the **confidence level**

Think ring toss…

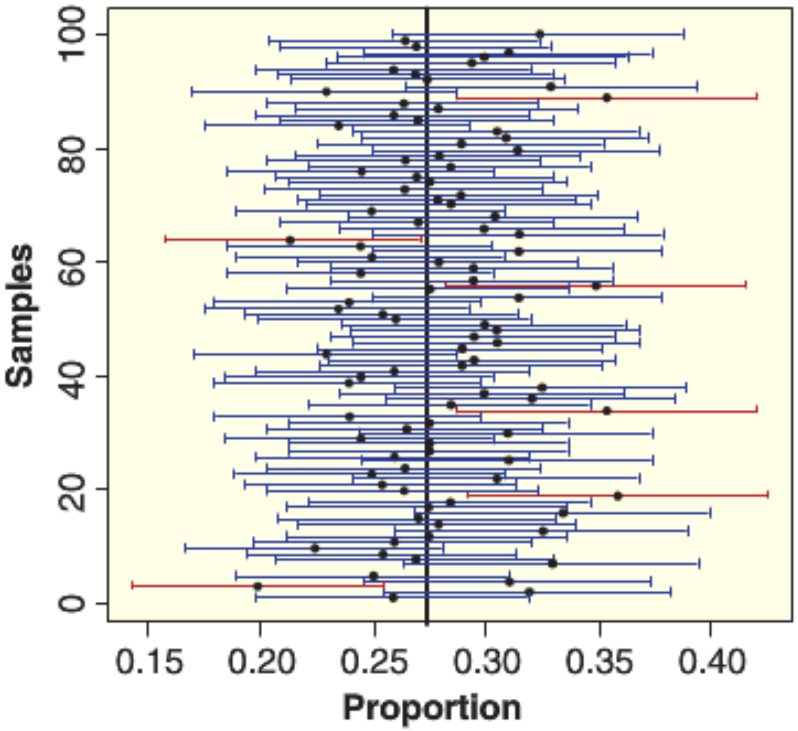
Parameter exists in the ideal world

We toss intervals at it

95% of those intervals capture the parameter

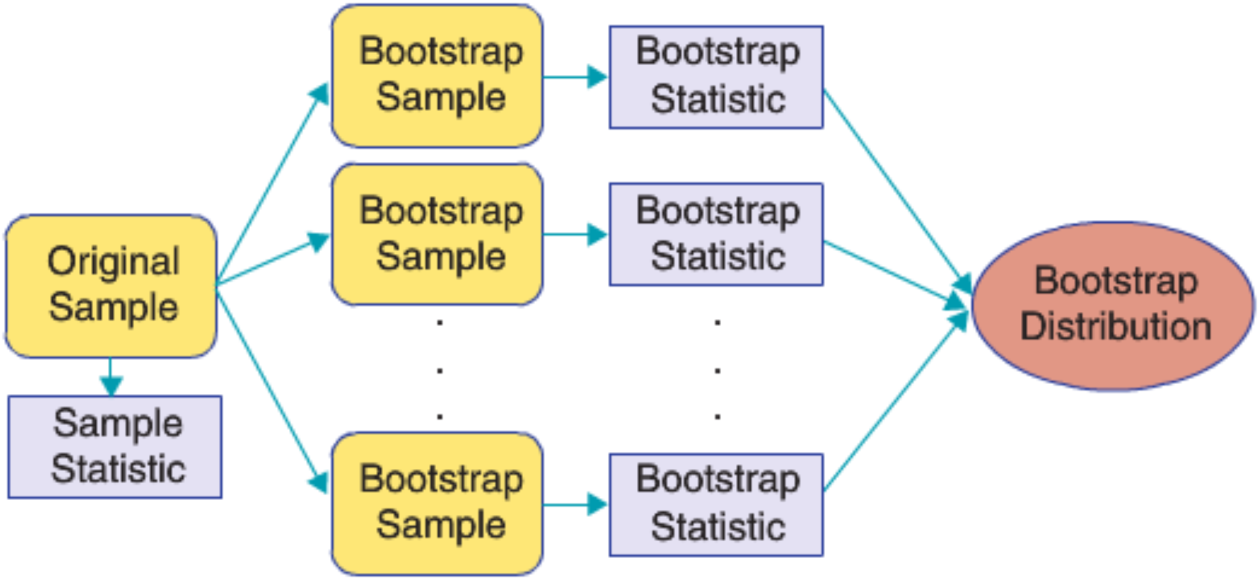
(for a 95% CI)

Wits and Wagers…



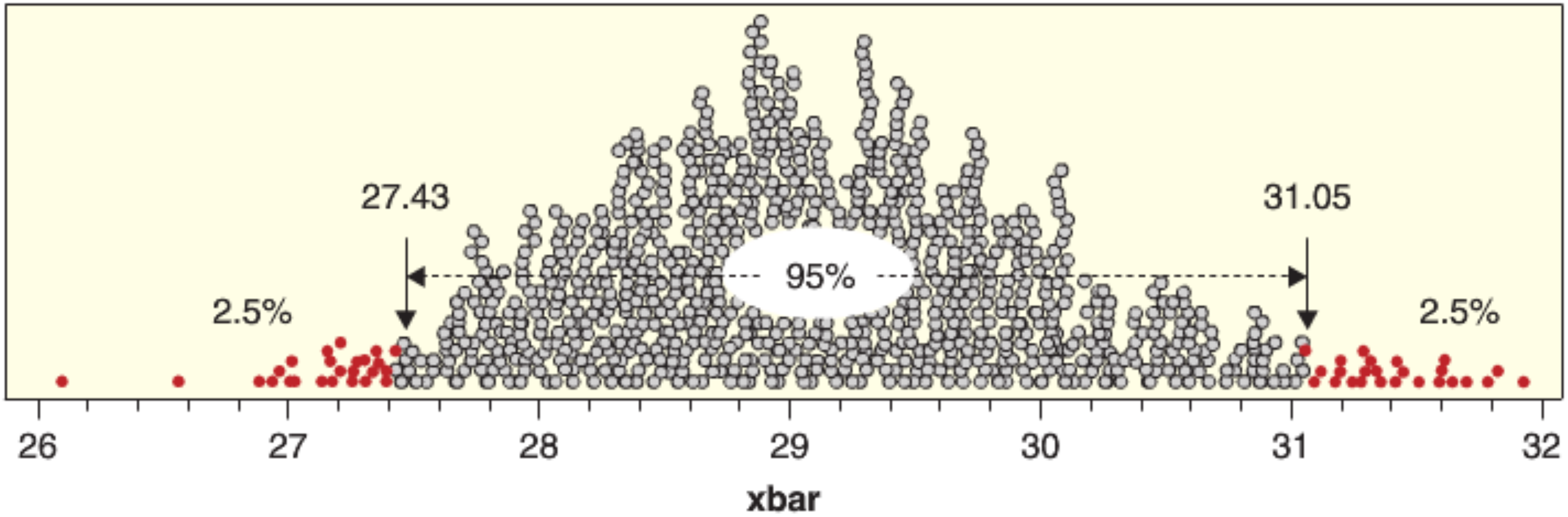
The bootstrap to estimate standard errors

The plug-in principle



95% Confidence Intervals

If the bootstrap distribution is ~symmetric we can use percentiles to calculate a 95% confidence interval

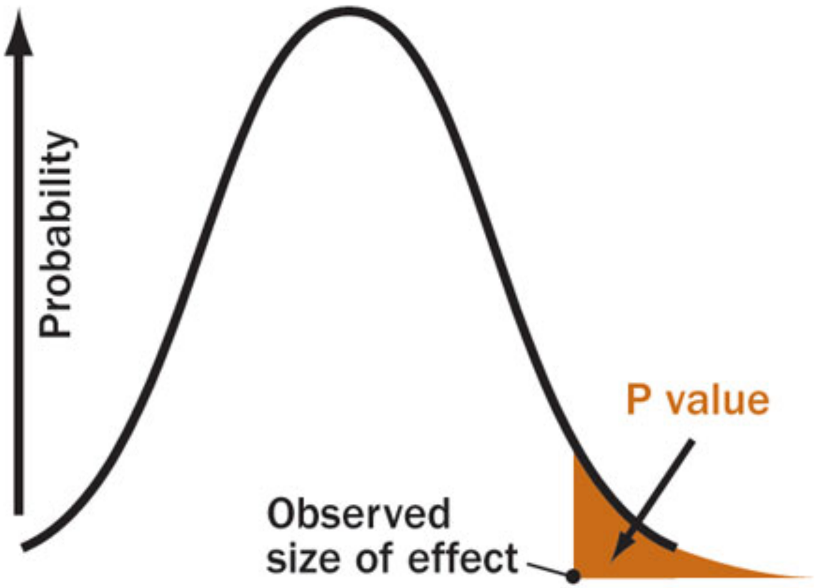


Formulas also exist for calculating CIs

- SE of the mean can be estimated as: $$SE = \frac{s}{\sqrt{n}}$$
- *CI is: Statistic ± 2 · SE*

Question: why is a p-value?

A p-value is the probability of getting a statistic as or more extreme than the observed statistic from a null model (null distribution)



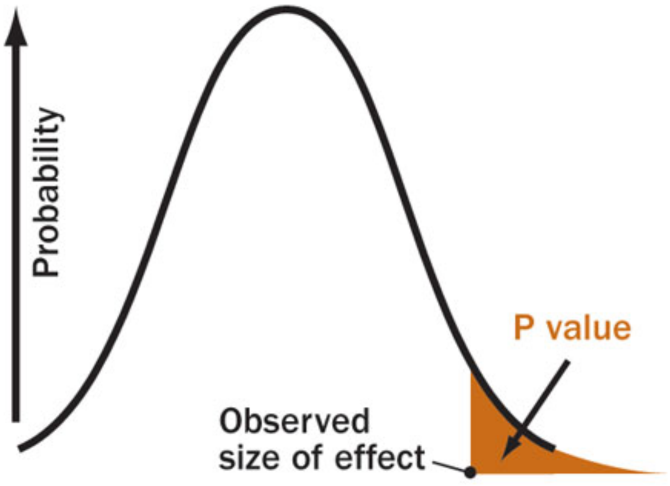
Hypothesis tests in 2 steps

1. Create a distribution showing what the statistic would like like if there was no effect (H_0 = true)

- Null distribution   (distribution if no effect)

2. Show that the statistic you observed is unlikely to come from this null distribution

- P-value is: $\Pr(S \geq \text{obs\_stat} \mid H_0)$
- P-value is **not**: $\Pr(H_0 \mid \text{data})$



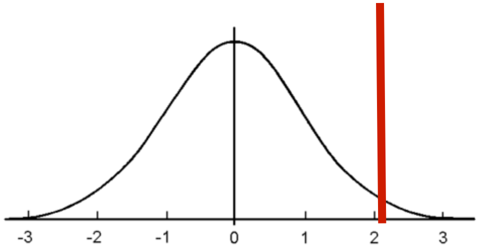
Trial metaphor of hypothesis testing

1. Assume innocence: H_0 is true
   - State H_0 and H_A

2. Gather evidence
   - Calculate the observed statistic

3. Create a distribution of what evidence would look like if H_0 is true
   - Null distribution

4. Assess the probability that the observed evidence would come from the null distribution
   - p-value



5. Make a judgement
   - Assess whether the results are statistically significant

Types of hypothesis tests test

Permutation test: Create null distribution simulating many random assignments

- 1. Shuffle the condition labels
- 2. Compute statistic on shuffled labels
- 3. Repeat many times to get a null distribution

Parametric tests: based on the fact that distribution of statistic is known

- t-tests, ANOVAs, chi-squared tests, etc...
- Less robust (based on Normal approximations)

Visual hypothesis tests

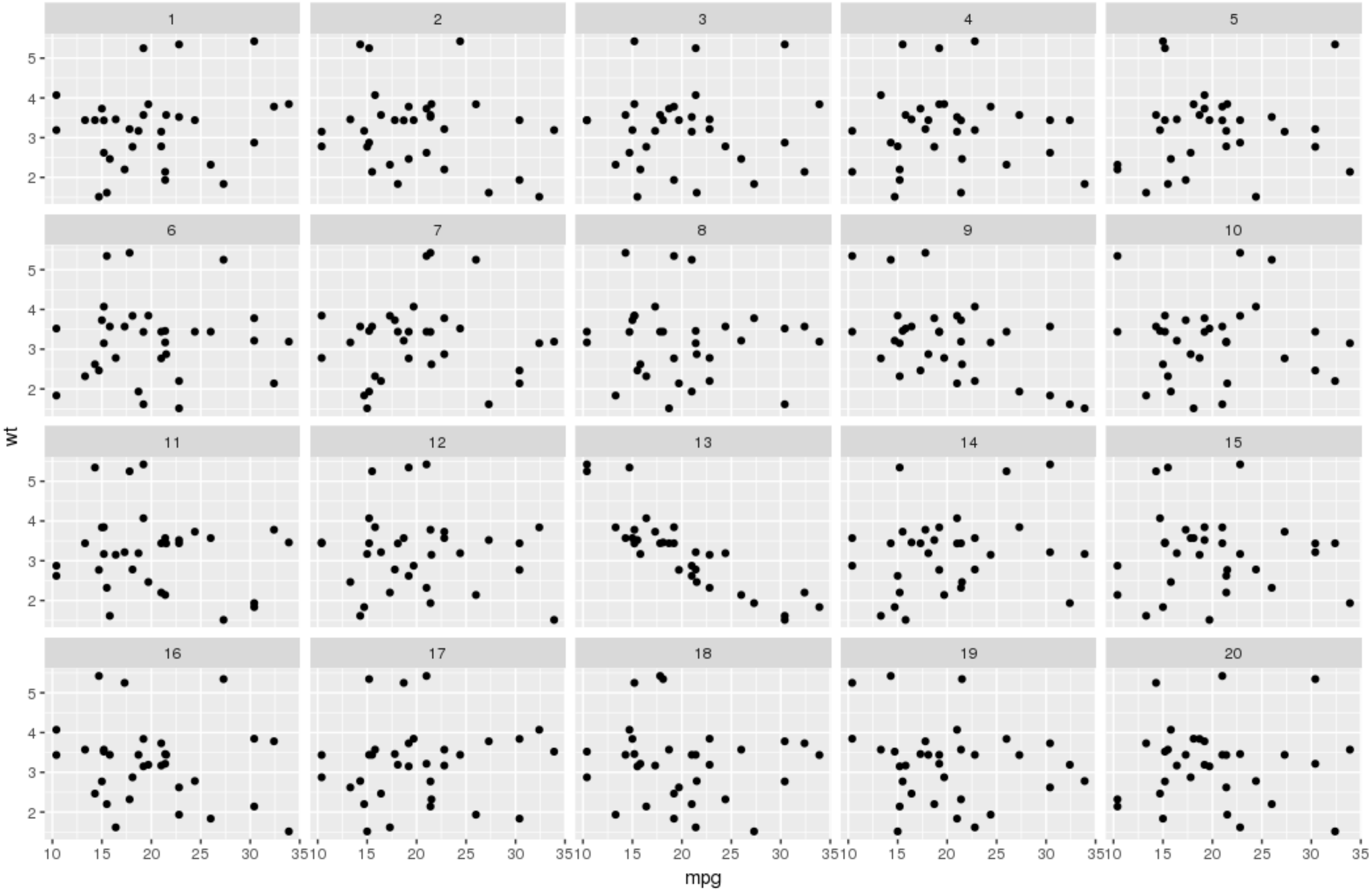
“Visual hypothesis tests” create visual plots of randomly rearranged points

- i.e., a set of ‘innocent’ plots where there is no pattern

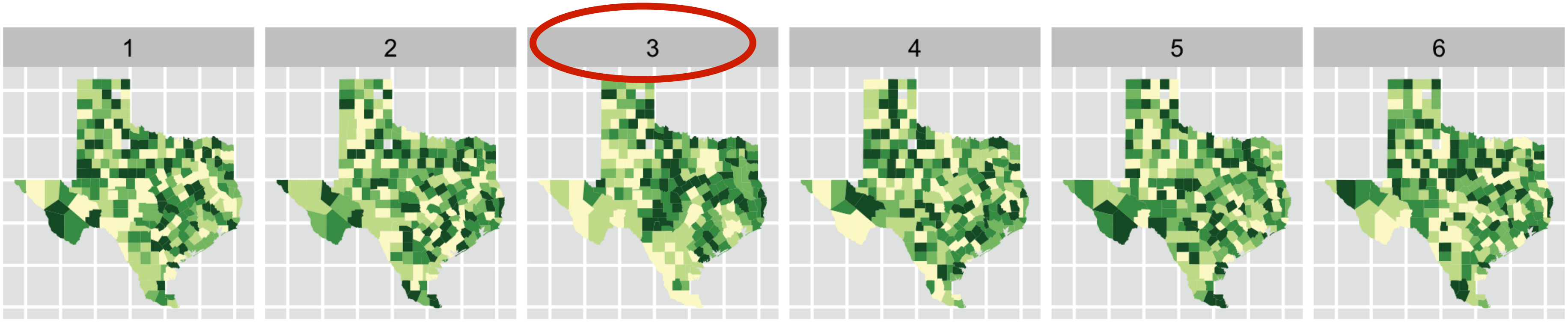
If you can tell the real data from the plots of randomly generated data this is equivalent to rejecting the null hypothesis

Visual hypothesis tests

Which plot shows the real data?



Visual hypothesis tests



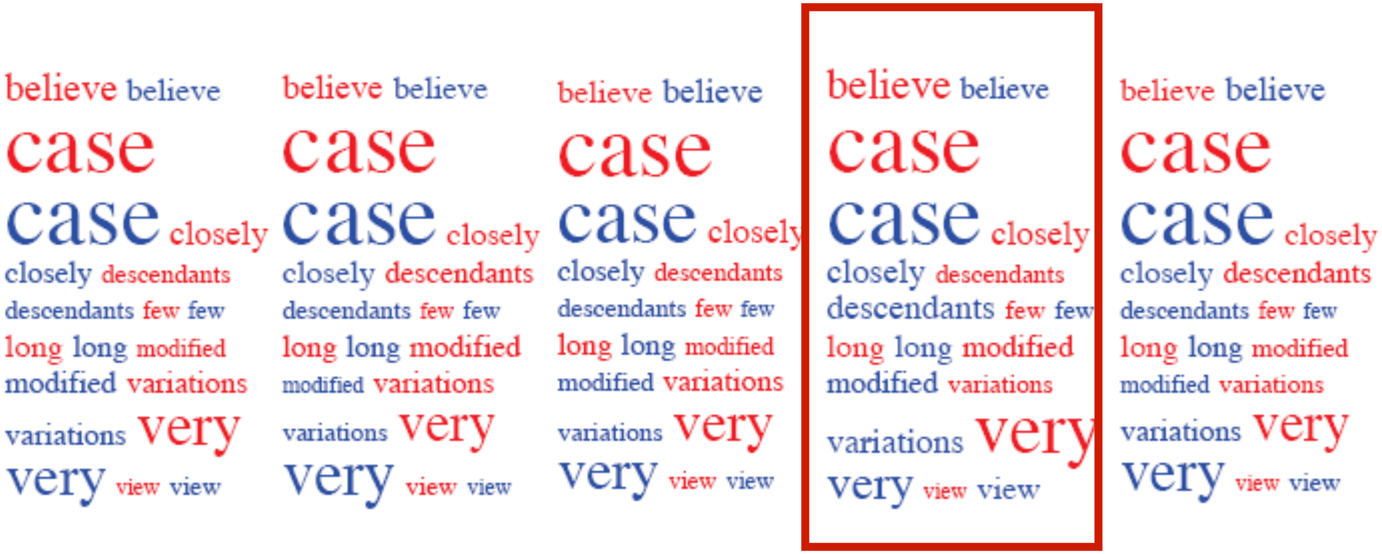
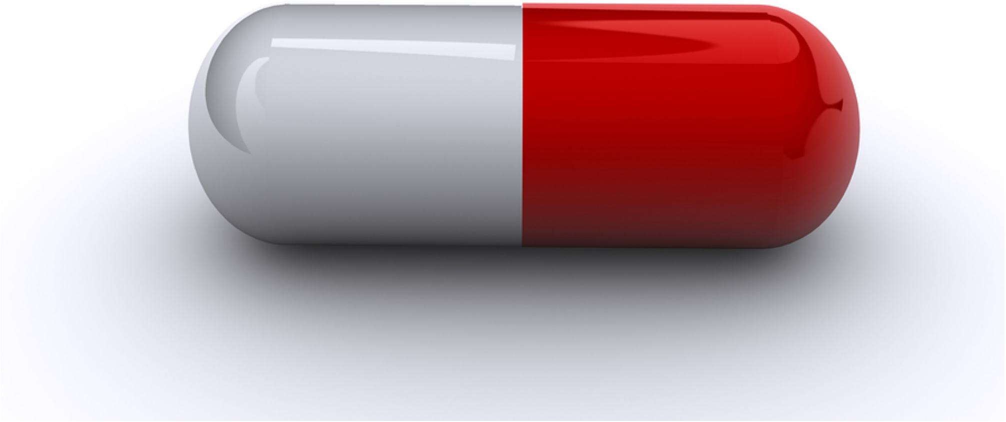


Fig. 5. Five tag clouds of selected words from the 1st (red) and 6th (blue) editions of Darwin's "Origin of Species". Four of the tag clouds were generated under the null hypothesis of no difference between editions, and one is the true data. Can you spot it?

See Wickham, et al, 2010

Permutation test example: Hypothesis tests for comparing two means

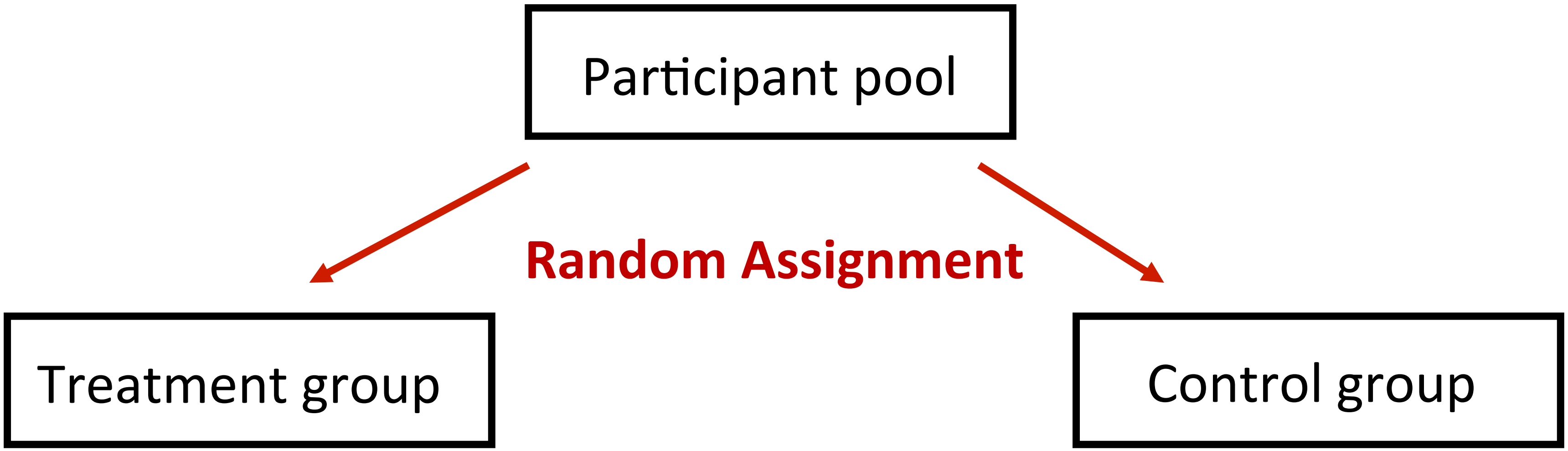


Question: Is this pill effective?

Experimental design

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get the pill
- Half in a *control group* where they get a fake pill (placebo)
- See if there is more improvement in the treatment group compared to the control group



Participant pool

Random Assignment

Treatment group

Control group

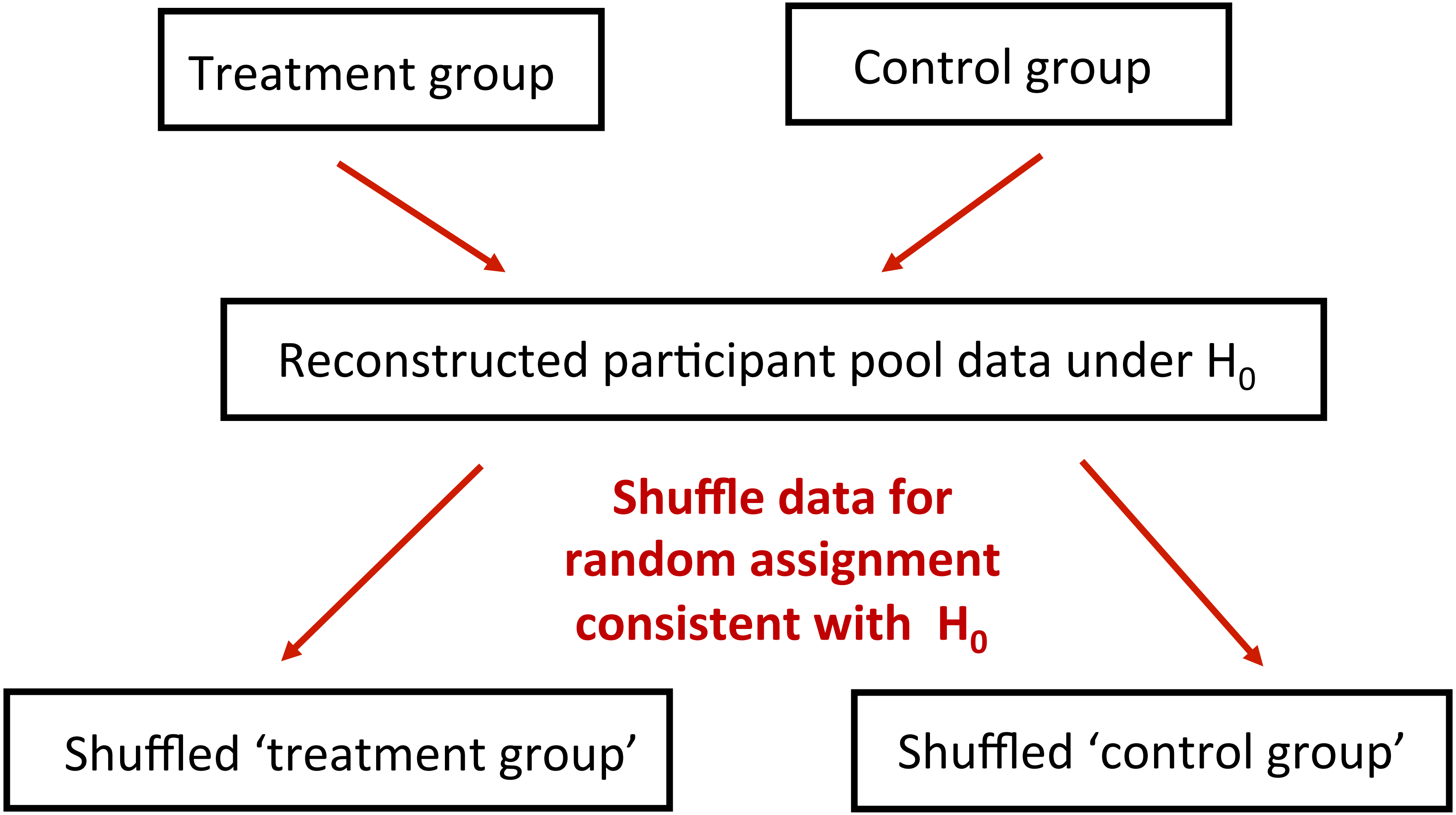
Hypothesis tests for differences in two group means

1. State the null and alternative hypothesis
    - H_0: $\mu_{Treatment} = \mu_{Control}$ or $\mu_{Treatment} – \mu_{Control} = 0$
    - H_A: $\mu_{Treatment} > \mu_{Control}$ or $\mu_{Treatment} – \mu_{Control} > 0$

2. Calculate statistic of interest
    - $\bar{x}_{Treatment} - \bar{x}_{Control}$

3. What should we do next…?

3. Create the null distribution!



Treatment group

Control group

Reconstructed participant pool data under H_0

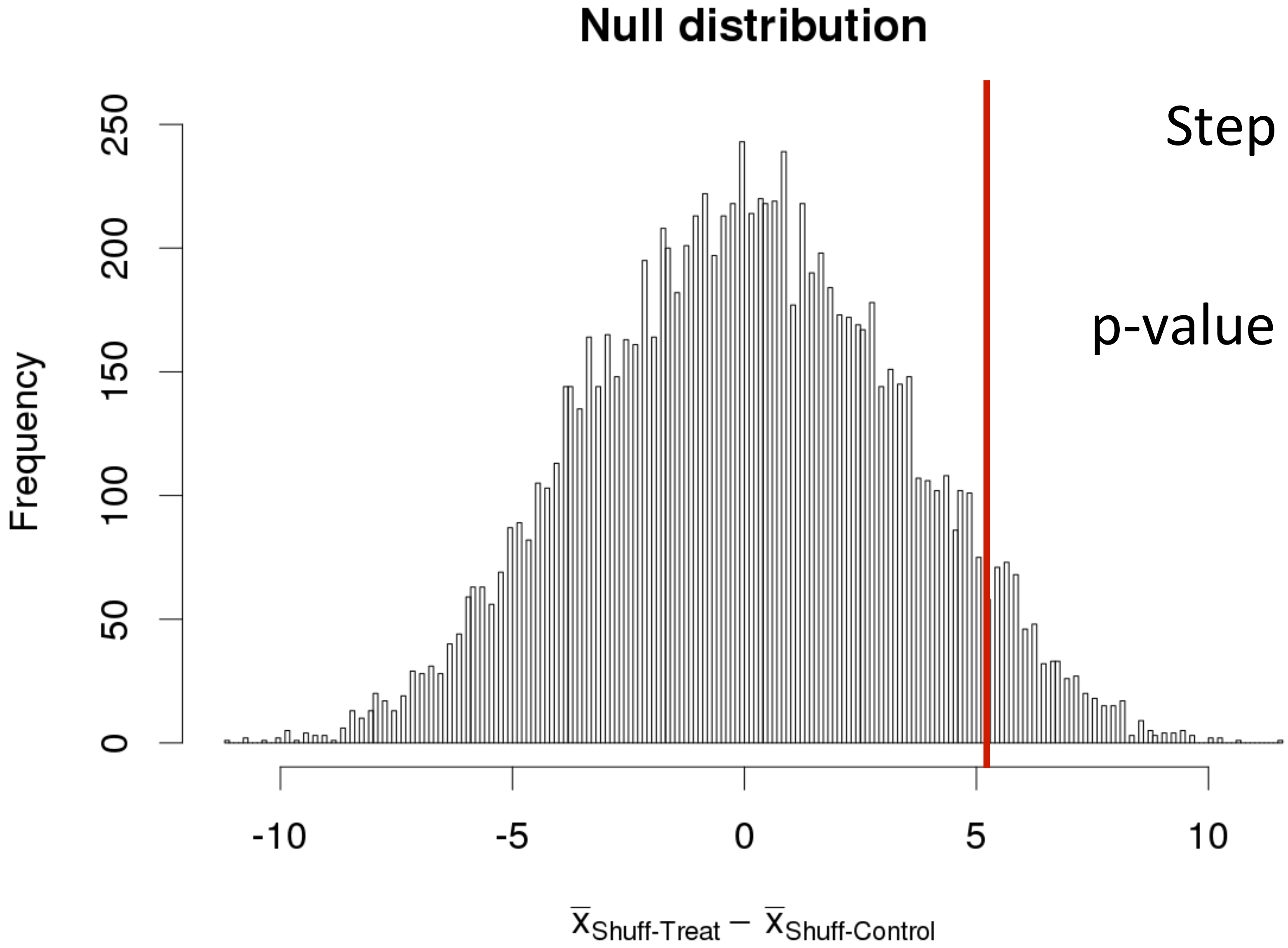
Shuffle data for random assignment consistent with H_0

Shuffled ‘treatment group’

Shuffled ‘control group’

One null distribution statistic: $\overline{x}_{Shuff\_Treatment} - \overline{x}_{Shuff\_control}$

Repeat 10,000 times



Step 4?

p-value = .064

Fisher vs. Neyman

Question: should you report exact p-values?

- p < .05
- p = .021

Null-hypothesis significance testing (NHST) is a hybrid of two theories:

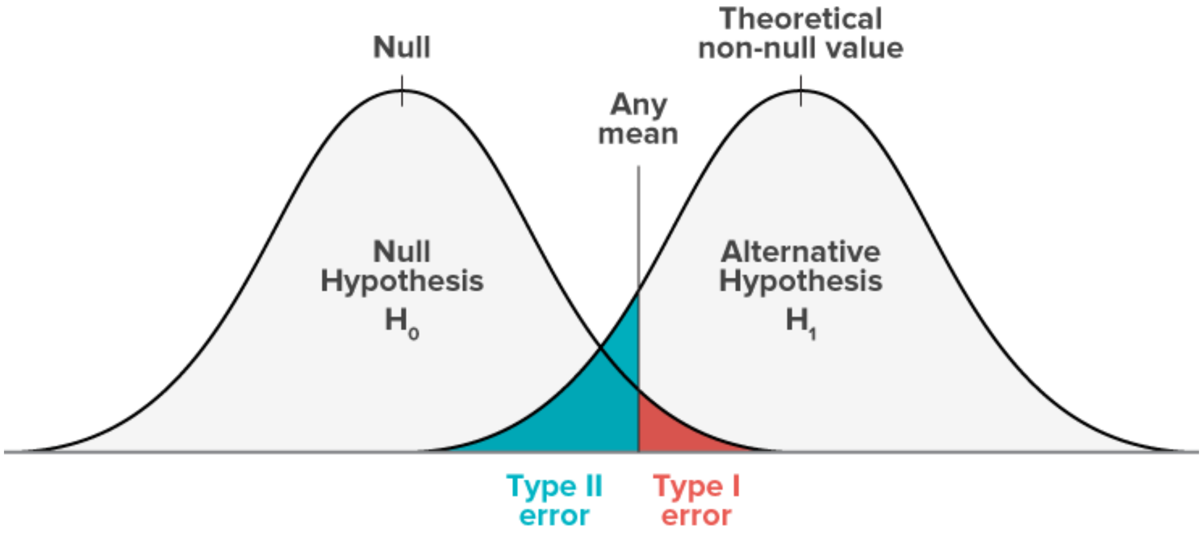
- 1. Significance testing of Ronald Fisher
- 2. Hypothesis testing of Jezy Neyman and Egon Pearson

Fisher (1890-1962)

Neyman (1894-1981)

Neat fact:
they hated
each other

Type I and Type II Errors

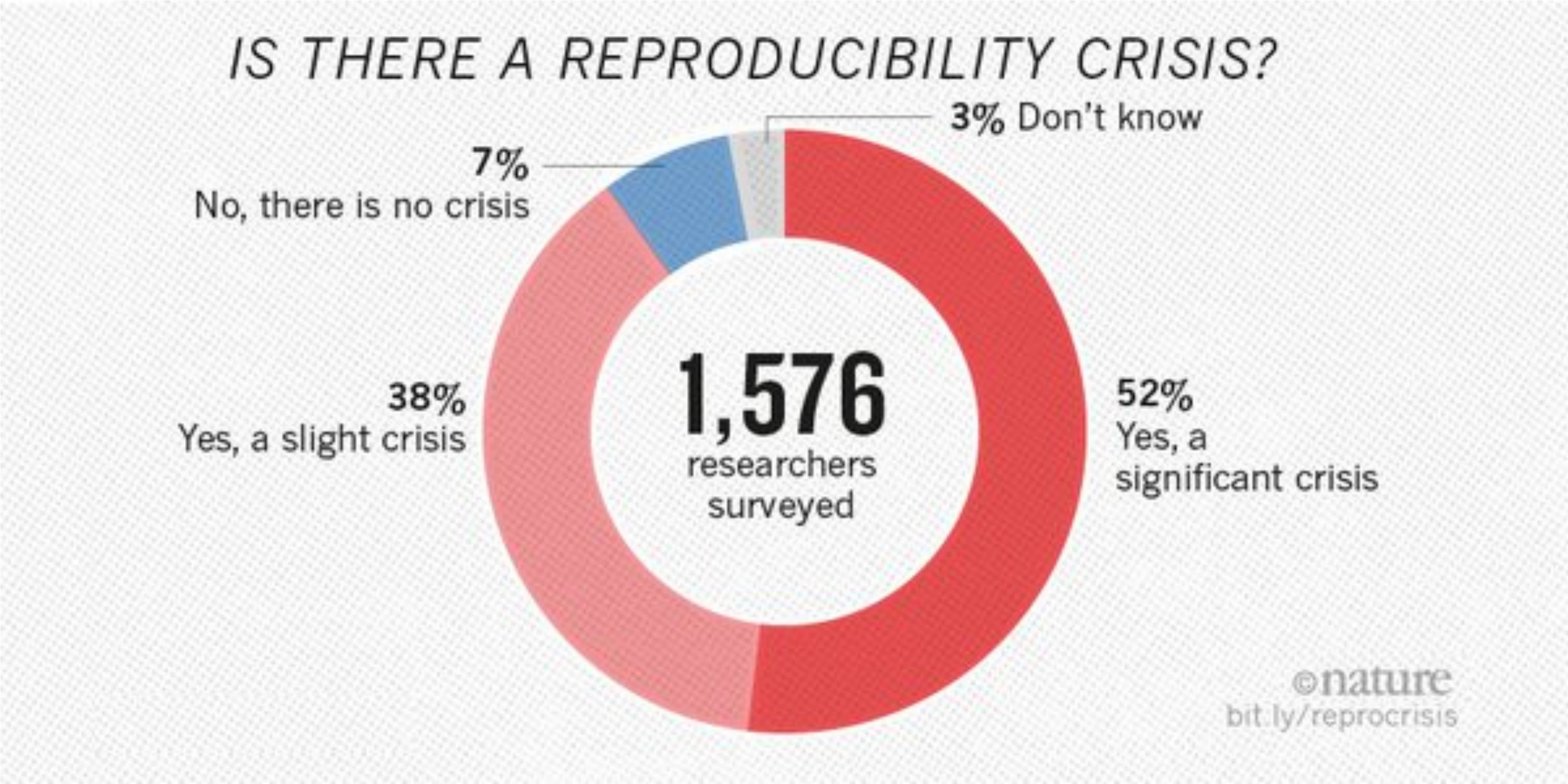


	Reject H_0	Do not reject H_0
H_0 is true	Type I error (α) (false positive)	No error
H_0 is false	No error	Type II error (β) (false negative)

Essay

Why Most Published Research Findings Are False

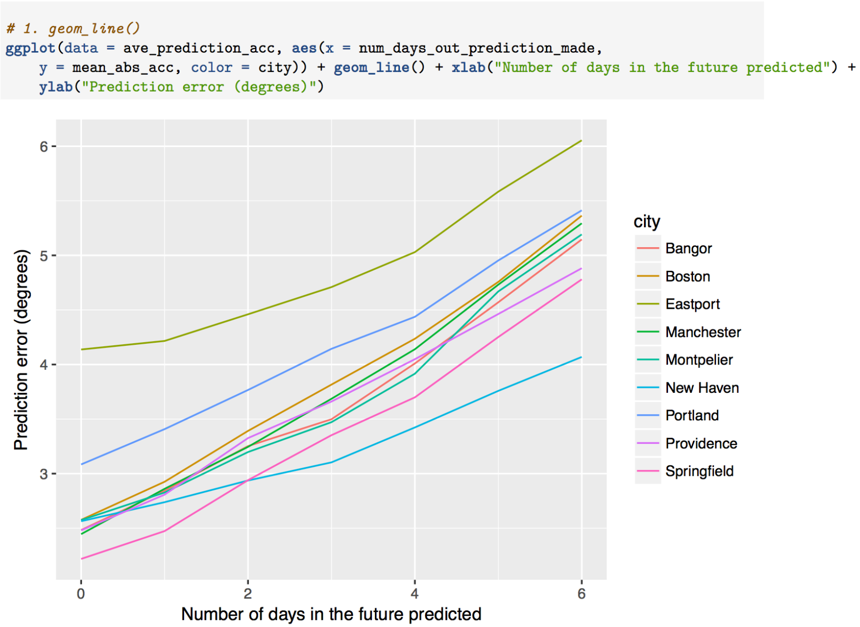
John P. A. Ioannidis



Do reproducible research!

Tools exist that allow you to embed your analyses into your written reports:

- R: R Markdown
- Python/Julia/R: Jupyter
- MATLAB: Live Scripts



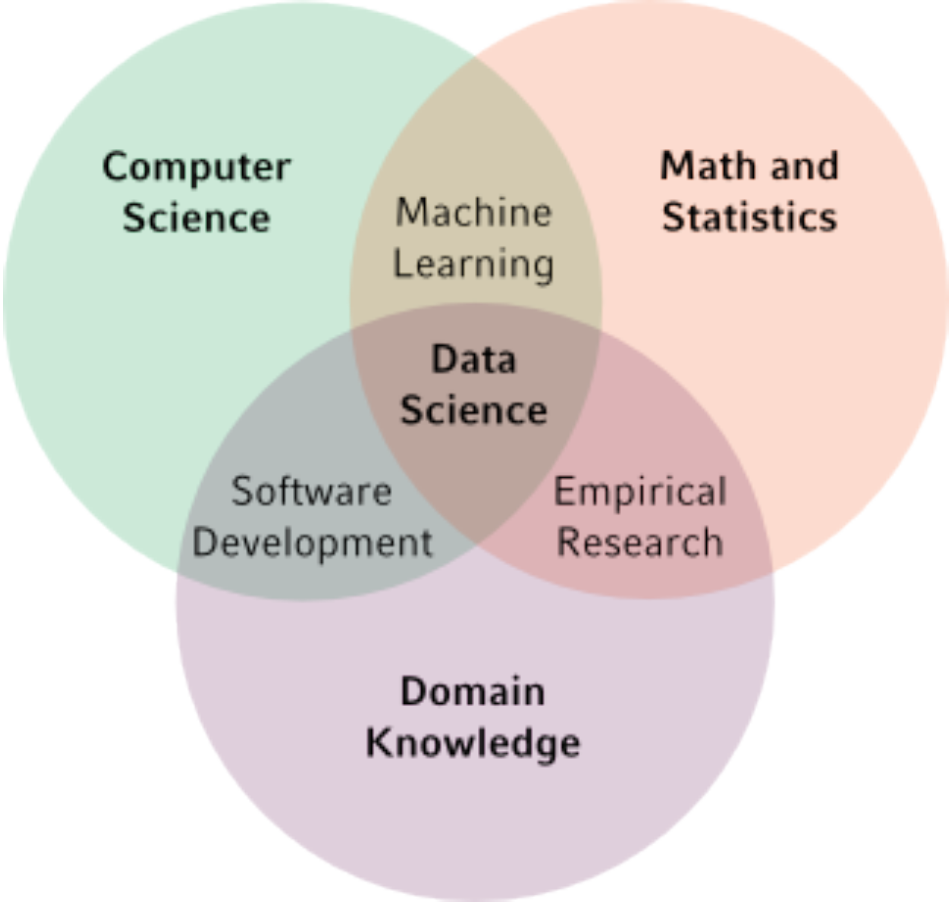
```
# 1. geom_line()
ggplot(data = ave_prediction_acc, aes(x = num_days_out_prediction_made,
    y = mean_abs_acc, color = city)) + geom_line() + xlab("Number of days in the future predicted") +
    ylab("Prediction error (degrees)")
```

John Tukey

“Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question”

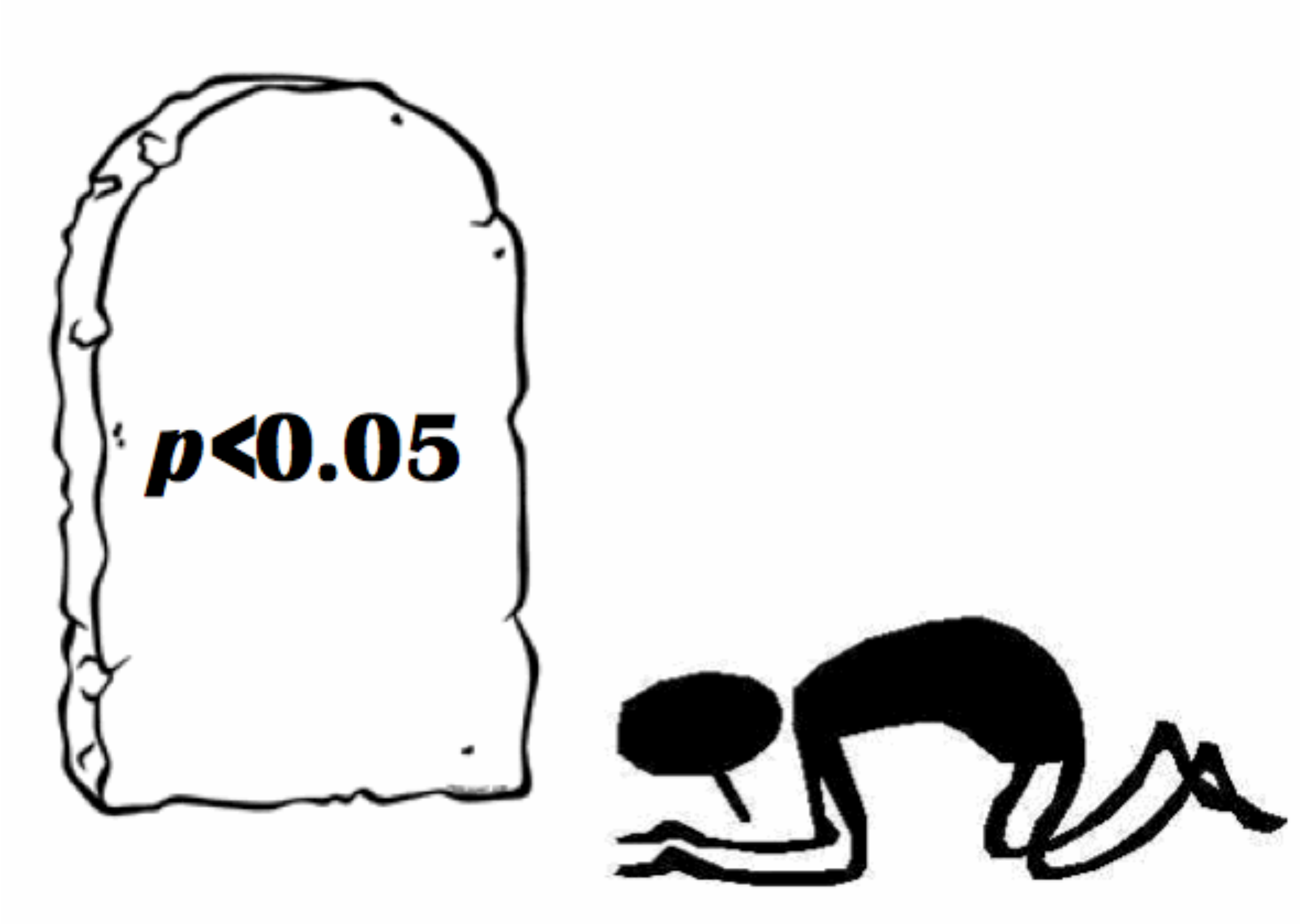
- *The future of data analysis*. Annals of Mathematical Statistics 33 (1), (1962), page 13.

What is Data Science?



See Donoho, 2017

Questions

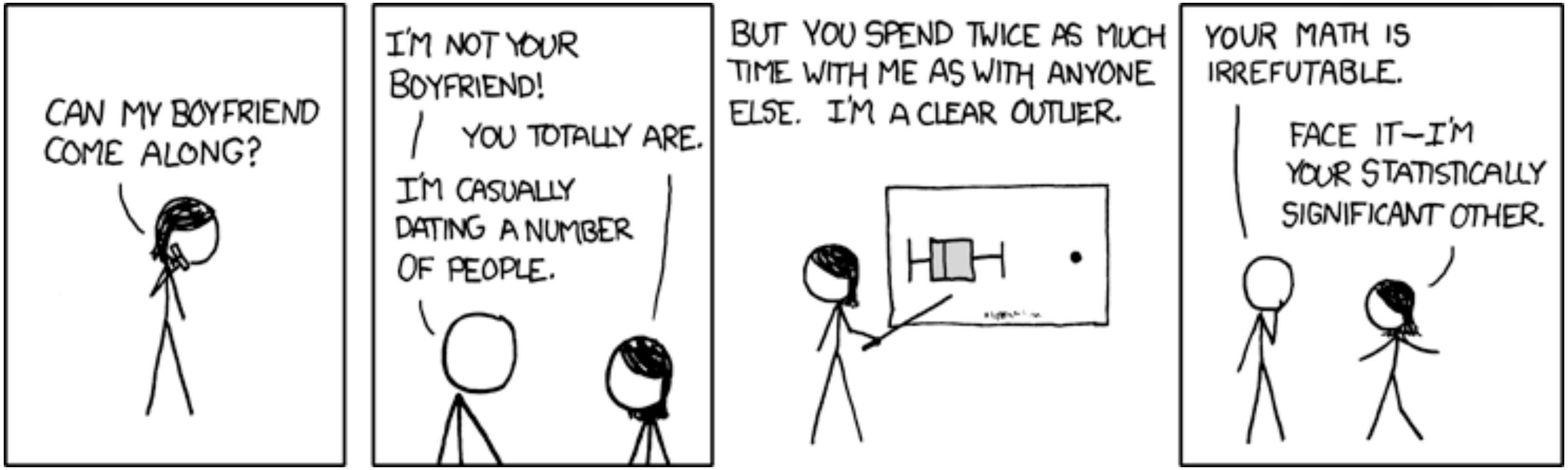


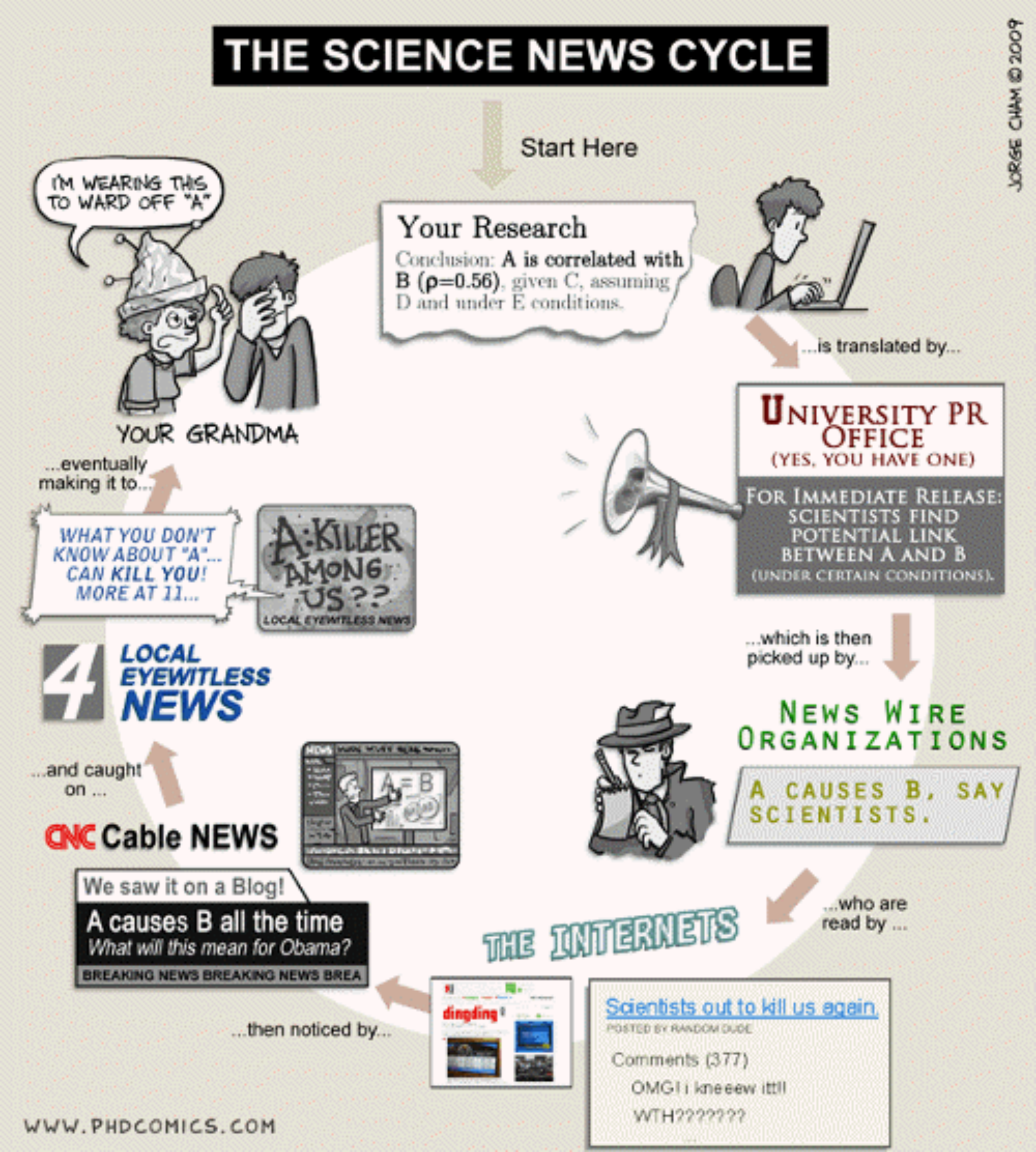
Interactive single neuron analysis tutorial…

https://neuraldata.net/spike/

Created by Brooke FItzgerald

Are we feeling ok about hypothesis tests?

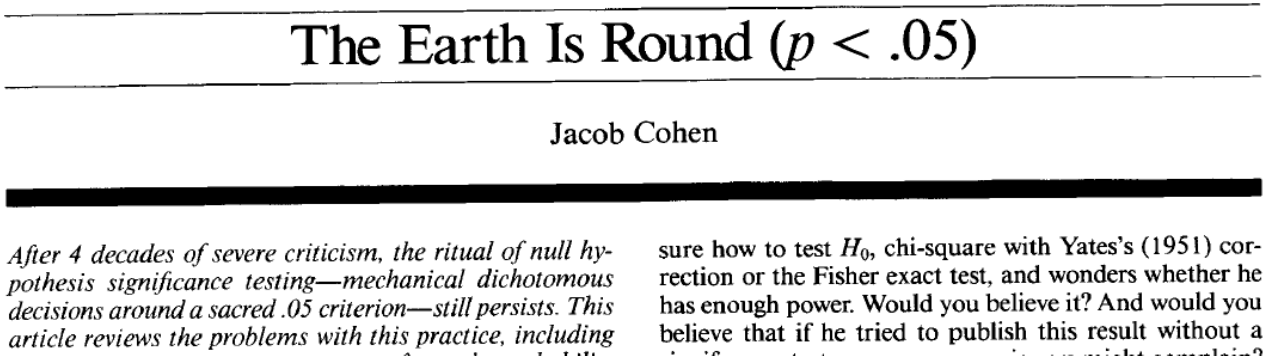

CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.


THE SCIENCE NEWS CYCLE
JORGE CHAM © 2009
Start Here
Your Research
Conclusion: A is correlated with B (ρ=0.56), given C, assuming D and under E conditions.
...is translated by...
UNIVERSITY PR OFFICE
(YES, YOU HAVE ONE)
FOR IMMEDIATE RELEASE: SCIENTISTS FIND POTENTIAL LINK BETWEEN A AND B (UNDER CERTAIN CONDITIONS).
...which is then picked up by...
NEWS WIRE ORGANIZATIONS
A CAUSES B, SAY SCIENTISTS.
...who are read by ...
THE INTERNETS
dingding
Scientists out to kill us again.
POSTED BY RANDOM DUDE
Comments (377)
OMG! i kneeew it!!
WTH???????
...then noticed by...
CNC Cable NEWS
We saw it on a Blog!
A causes B all the time
What will this mean for Obama?
BREAKING NEWS BREAKING NEWS BREA
A = B
...and caught on ...
4 LOCAL EYEWITLESS NEWS
WHAT YOU DON'T KNOW ABOUT "A"... CAN KILL YOU! MORE AT 11...
A: KILLER AMONG US??
LOCAL EYEWITLESS NEWS
...eventually making it to...
YOUR GRANDMA
I'M WEARING THIS TO WARD OFF "A"
WWW.PHDCOMICS.COM

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis



The Earth Is Round ($p < .05$)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including

sure how to test H_0, chi-square with Yates's (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it? And would you believe that if he tried to publish this result without a

American Statistical Association's Statement on p-values