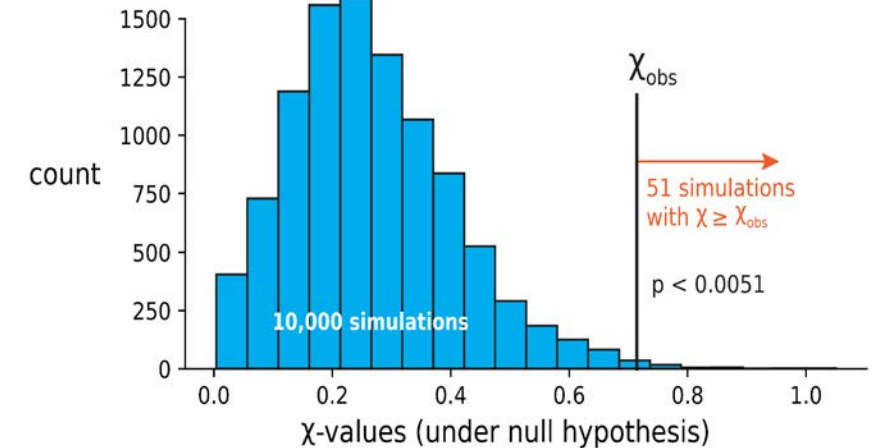
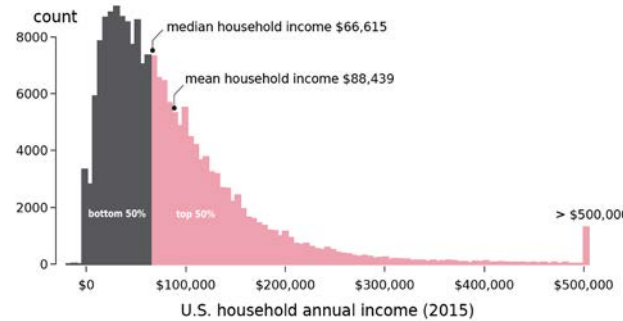
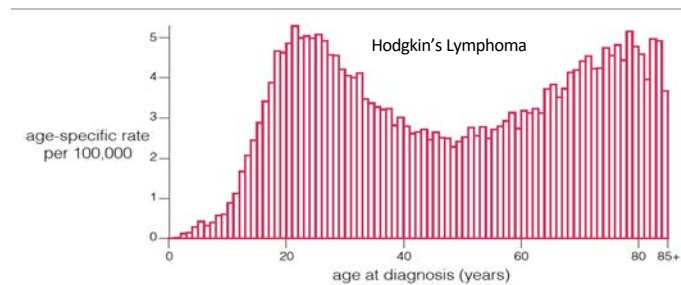
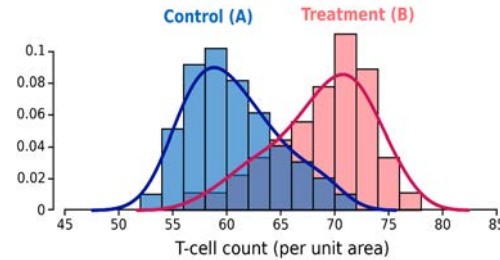
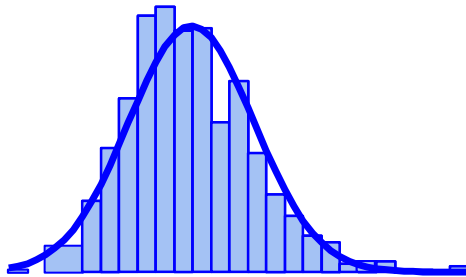


Introductory Statistics through Resampling: Addressing the Reproducibility Crisis in the Sciences



Alan Garfinkel

**Professor of Medicine (Cardiology) and Integrative Biology and Physiology
UCLA**

Harvard Workshop
Fri 7/30 5:15-6:15

“Statistical Significance”

The purpose of a standard statistics course is to teach you how to determine if a result is
“statistically significant”

The formulas for the t-test, χ^2 , etc. are formulas for determining statistical significance

What is a statistics course?

- memorize formulas
- input your data into the formula to get a number
- plug that number into another even more mysterious formula, which is hidden from you
- this magic formula tells you whether or not your data is “statistically significant”

t-test

$$t = \sqrt{n} \left(\frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{AB}} \right)$$

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2}$$

note: requires

1. mean values
2. algebraic difference
3. standard deviation

magic formula

$$p(t \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$
$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$$

$\nu = n - 1$

If $t > 1.96$ then $p < 0.05$
“statistically significant”

What is a statistics course?

- memorize formulas
- input your data into the formula to get a number
- plug that number into another even more mysterious formula, which is hidden from you
- this magic formula tells you whether or not your data is “statistically significant”

χ^2 -test

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad \chi^2 = \frac{1}{2^{d/2} \Gamma(d/2)} x^{d/2-1} e^{-x/2}$$

$$\text{where } \Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

If $\chi^2 > 3.84$ then $p < 0.05$ “statistically significant”

You plug your numbers into a magic formula, and the formula tells you whether your result is “**statistically significant**”

- Is this good pedagogy?
- “shut up and calculate”



- Is this good science?

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

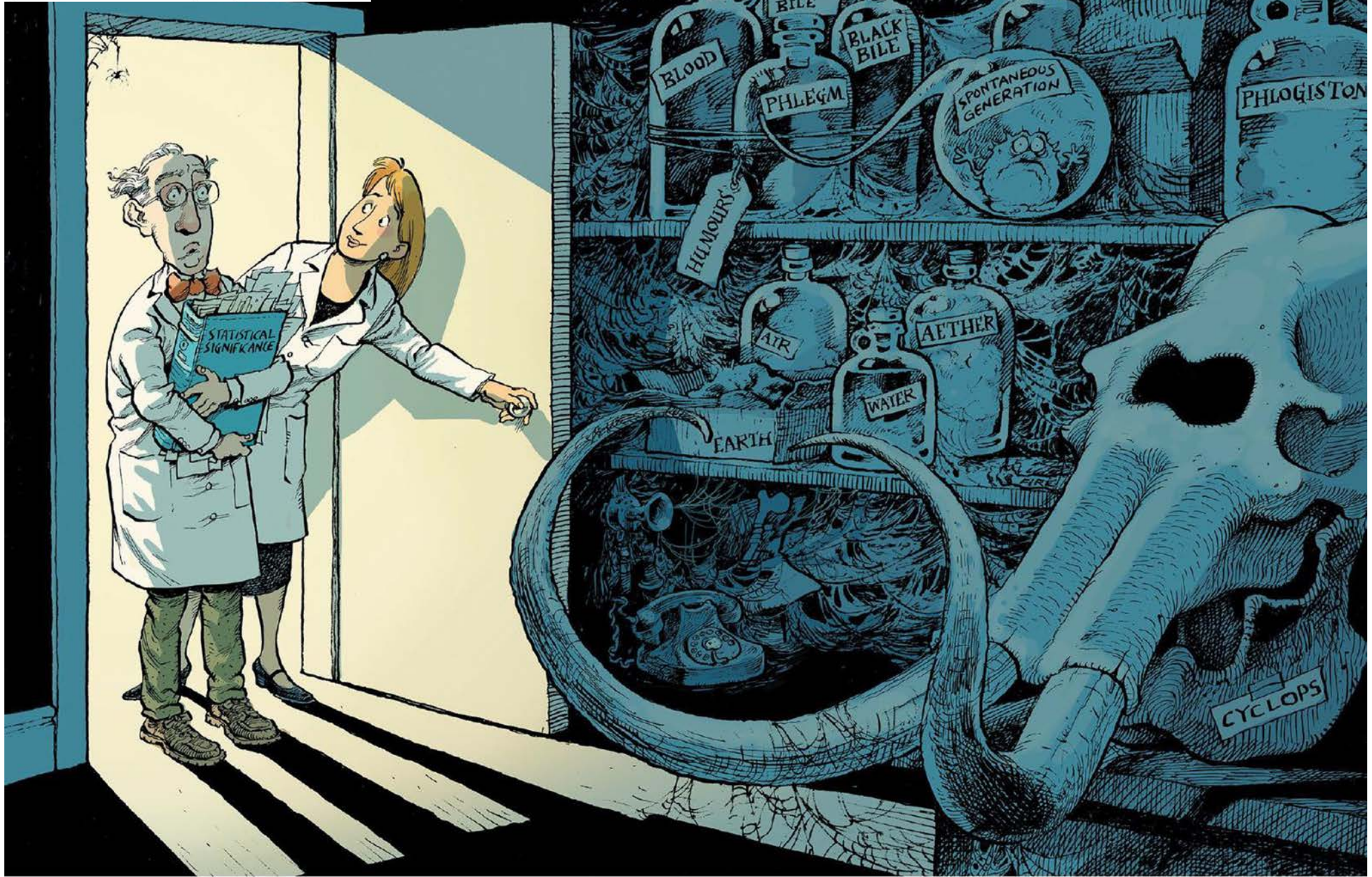


“We ... call for the entire concept of statistical significance to be abandoned.”

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

21 MARCH 2019 | VOL 567 | NATURE | 305



**Why “abandon statistical
significance?”**

Is there a problem?

yes!

In science, irreproducible research is a quiet crisis

By Carolyn Johnson Globe staff, March 19, 2015, 3:46 p.m.



Attempts to quantify how much research is wrong have been sobering.

- Scientists at Bayer HealthCare reported that when they tried to reproduce 67 published discoveries in oncology, women's health, and cardiovascular disease, only a quarter of the time were their in-house results completely consistent with what had been reported.
- A scientist at Amgen disclosed that, despite concerted effort, the company had successfully repeated only six of 53 major cancer findings.
- Last year, neuroscience researchers tried to repeat five studies that showed 17 links between certain brain structures and behaviors — for example, a finding that people with more gray matter in particular brain regions also have more Facebook friends. The researchers could not replicate any of the connections and found evidence for what scientists call the “null hypothesis” — the idea that there was no relationship.

Newsweek

Why Almost Everything You Hear About Medicine Is Wrong

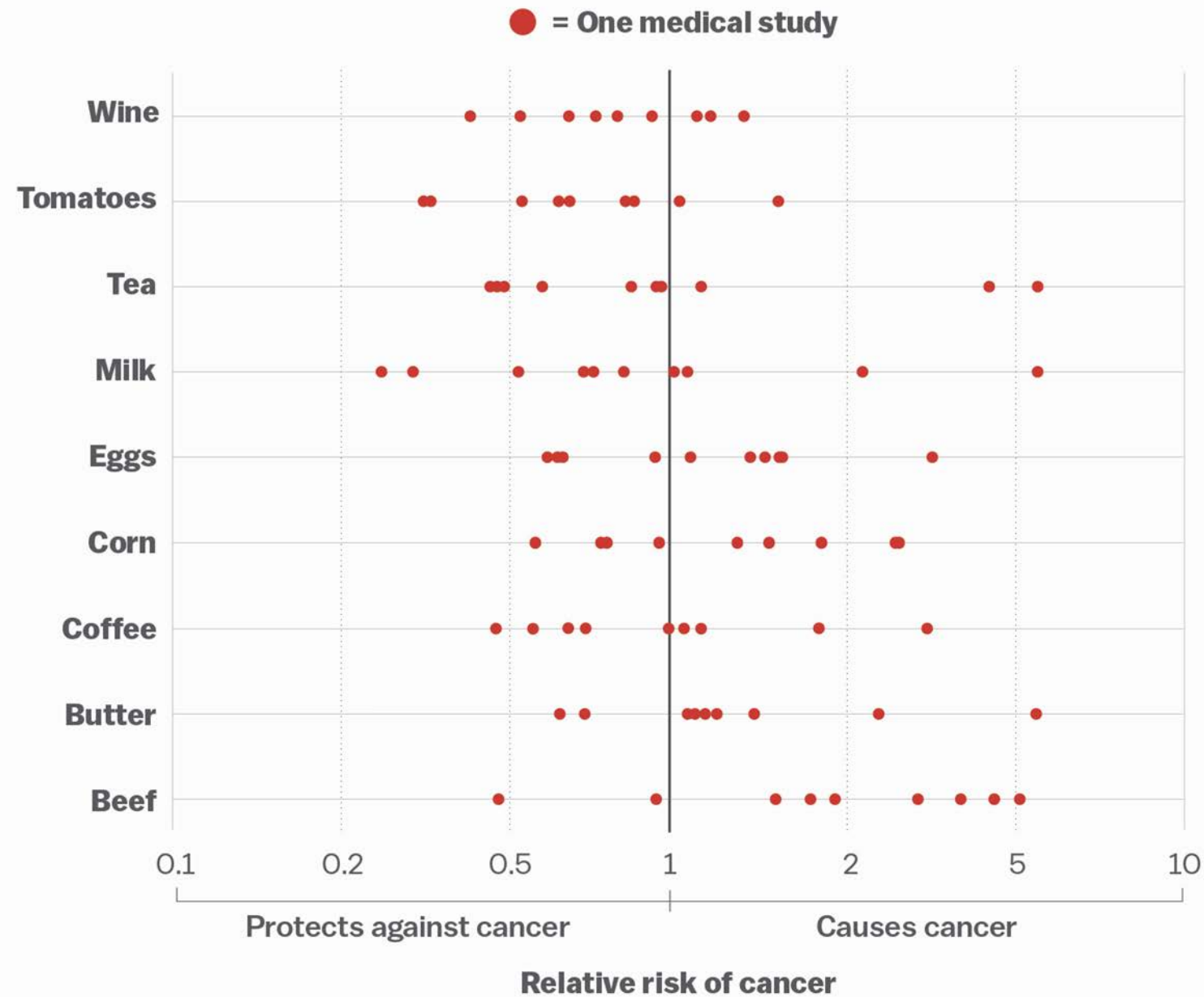


by [Sharon Begley \(/authors/sharon-begley.html\)](https://www.newsweek.com/authors/sharon-begley.html) January 24, 2011



Illustration by Jacob Thomas

But what if wrong answers aren't the exception but the rule?



In each case, the finding was “statistically significant”

ESSAY | ODDS ARE, IT'S WRONG

Odds Are, It's Wrong

Science fails to face the shortcomings of statistics

By Tom Slegfried

Science News 2010

Nowhere are the problems with statistics more blatant than in studies of genetic influences on disease. In 2007, for instance, researchers combing the medical literature found numerous studies linking a total of 85 genetic variants in 70 different genes to acute coronary syndrome, a cluster of heart problems. When the researchers compared genetic tests of 811 patients that had the syndrome with a group of 650 (matched for sex and age) that didn't, only one of the suspect gene variants turned up substantially more often in those with the syndrome — a number to be expected by chance.



"Our null results provide no support for the hypothesis that any of the 85 genetic variants tested is a susceptibility factor" for the syndrome, the researchers reported in the *Journal of the American Medical Association*.

2017

Biological
Psychiatry

No Evidence That Schizophrenia Candidate Genes Are More Associated With Schizophrenia Than Noncandidate Genes

Emma C. Johnson, Richard Border, Whitney E. Melroy-Greif, Christiaan A. de Leeuw, Marissa A. Ehringer, and Matthew C. Keller

From the Department of Psychology and Neuroscience (ECJ, RB, MCK), Institute for Behavioral Genetics (ECJ, RB, MAE, MCK), and Department of Integrative Physiology (MAE), University of Colorado Boulder, Boulder, Colorado; Department of Neuroscience (WEM-G), The Scripps Research Institute, La Jolla, California; and the Department of Complex Trait Genetics (CAAdL), Centre for Neurogenomics and Cognitive Research/VU University Amsterdam, Amsterdam; and Institute for Computing and Information Sciences (CAAdL), Radboud University Nijmegen, Nijmegen, the Netherlands.

No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples

Richard Border, M.A., Emma C. Johnson, Ph.D., Luke M. Evans, Ph.D., Andrew Smolen, Ph.D., Noah Berley, Patrick F. Sullivan, M.D., Matthew C. Keller, Ph.D.

Conclusions: The study results do not support previous depression candidate gene findings, in which large genetic effects are frequently reported in samples orders of magnitude smaller than those examined here. Instead, the results suggest that early hypotheses about depression candidate genes were incorrect and that the large number of associations reported in the depression candidate gene literature are likely to be false positives.

Am J Psychiatry 2019; 176:376–387; doi: 10.1176/appi.ajp.2018.18070881

The Institute for Behavioral Genetics (Border, Johnson, Evans, Smolen, Keller), the Department of Psychology and Neuroscience (Border, Berley, Keller), the Department of Applied Mathematics (Border), and the Department of Ecology and Evolutionary Biology (Evans), University of Colorado Boulder, Boulder; the Department of Psychiatry, Washington University School of Medicine, St. Louis (Johnson); the Department of Genetics and Psychiatry, University of North Carolina at Chapel Hill (Sullivan); and the Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm (Sullivan).



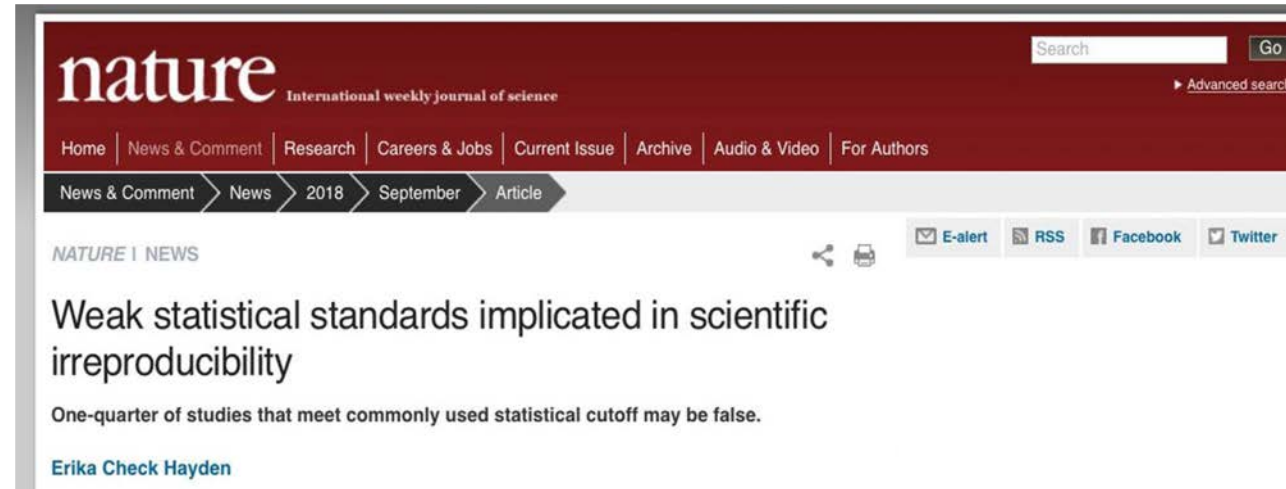
Performance of polygenic risk scores in screening, prediction, and risk stratification: secondary analysis of data in the Polygenic Score Catalog

Aroon D Hingorani,^{1,2,3,4} Jasmine Gratton ,^{1,2} Chris Finan,^{1,2,3,4} A Floriaan Schmidt,^{1,3,4,5} Riyaz Patel,^{1,2,3,4} Reecha Sofat,^{4,6} Valerie Kuan,^{1,2,3} Claudia Langenberg,^{7,8,9} Harry Hemingway,^{2,3,4,10} Joan K Morris,¹¹ Nicholas J Wald^{10,11}

CONCLUSION Polygenic risk scores performed poorly in population screening, individual risk prediction, and population risk stratification. Strong claims about the effect of polygenic risk scores on healthcare seem to be disproportionate to their performance.

What are the causes of this crisis?

#1- Bad statistics!





Commentary | Published: 26 February 2015

The fickle P value generates irreproducible results

Lewis G Halsey , Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

Nature Methods **12**, 179–185 (2015) | [Download Citation](#) 

The reliability and reproducibility of science are under scrutiny. However, a major cause of this lack of repeatability is not being considered: the wide sample-to-sample variability in the P value. We explain why P is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic.

What's wrong with the formulas?

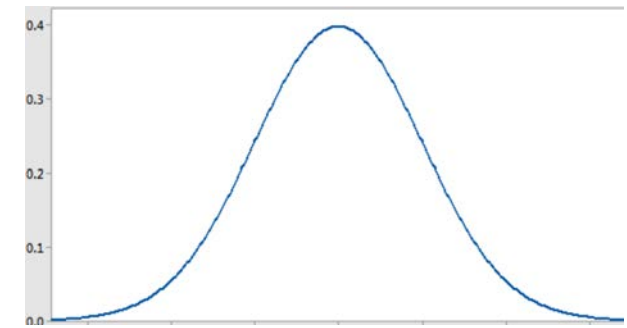
$$t = \frac{(\bar{a} - \bar{b})}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}}$$

if $t > 1.98$, then $p < .05$

$$p(t | \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Formulas are approximations based on heavy assumptions

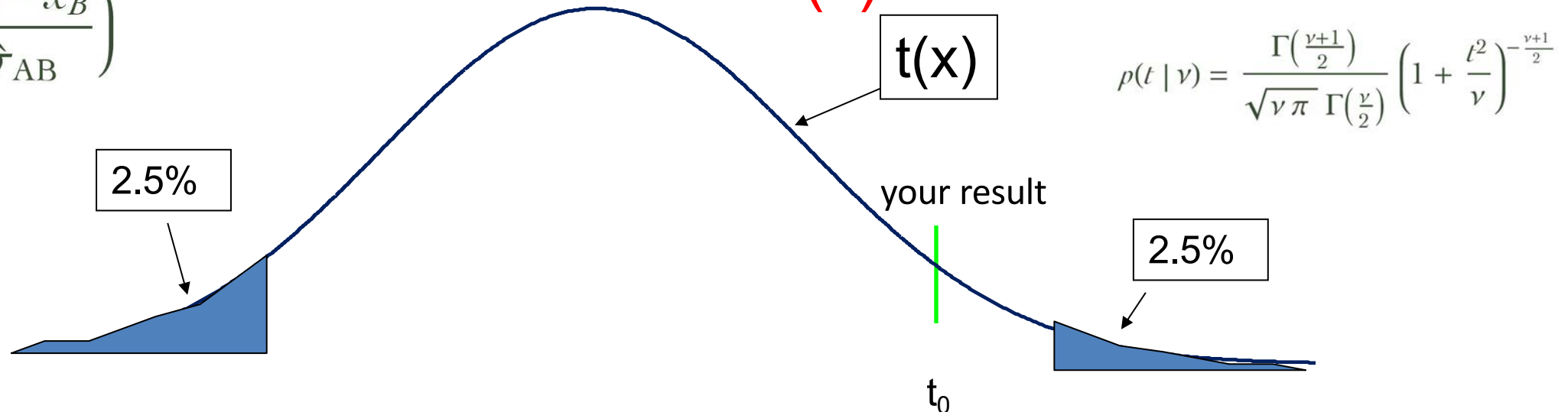
- normal distributions
- “as $n \rightarrow$ infinity”
- **must** use means and standard deviations



Theorem- Student's t-test

I**F** groups A and B are normally distributed (= bell-shaped)
a**n**d we use the mean ($\bar{A}$ and $\bar{B}$) to describe the groups
a**n**d we use subtraction to compare the two groups
a**n**d A and B have the same variance
then **t** will be distributed like a theoretical
distribution called “t(x)”

$$t = \sqrt{n} \left(\frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{AB}} \right)$$



old, formula-based statistics
is married to the Normal
distribution and “means and
standard deviations”

The correlation coefficient

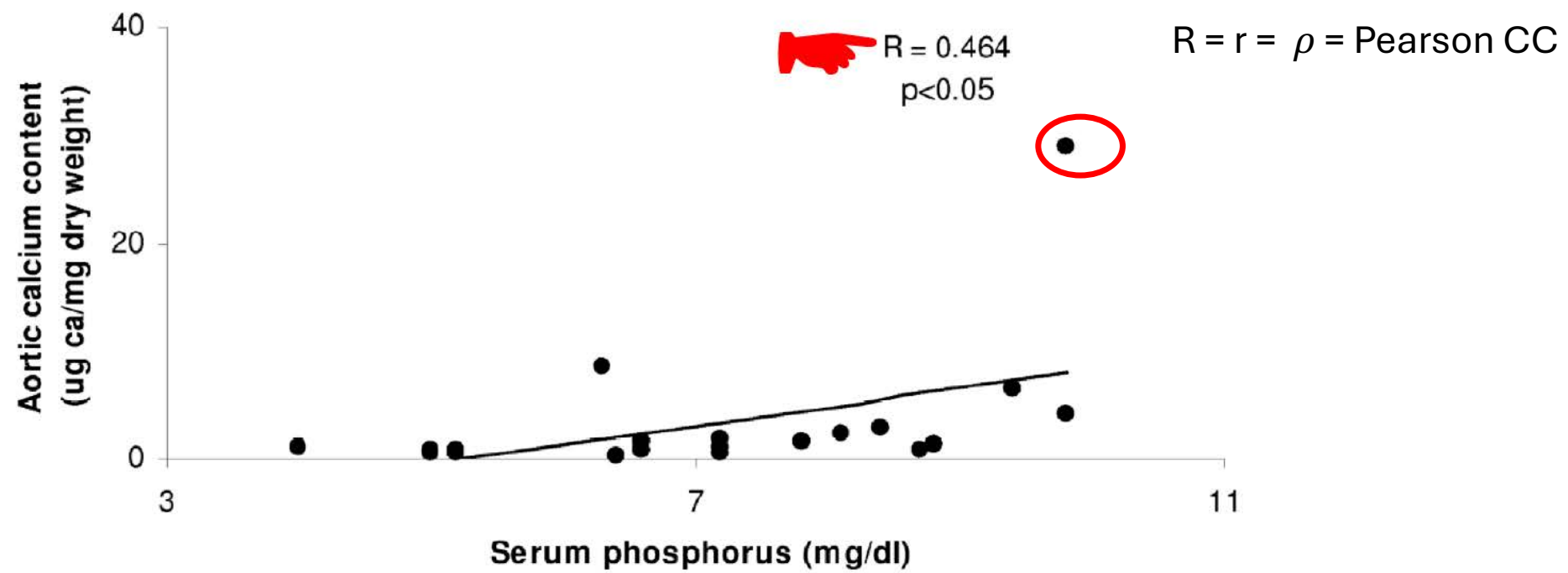
$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{Y})^2}}$$

**Note lots of means,
lots of squares**

The correlation coefficient is defined in terms of “means and standard deviations”

Assigning p-values: the theoretical calculation of the p-value of a given correlation coefficient depends heavily on key assumptions:

- X and Y must be jointly Normally distributed
- variance in X must be constant across Y, and variance in Y must be constant across X



Kidney International 2009

How Normal is the Normal Distribution?

Average tax refunds up \$40 for 2019



By Donna Borak, CNN Business

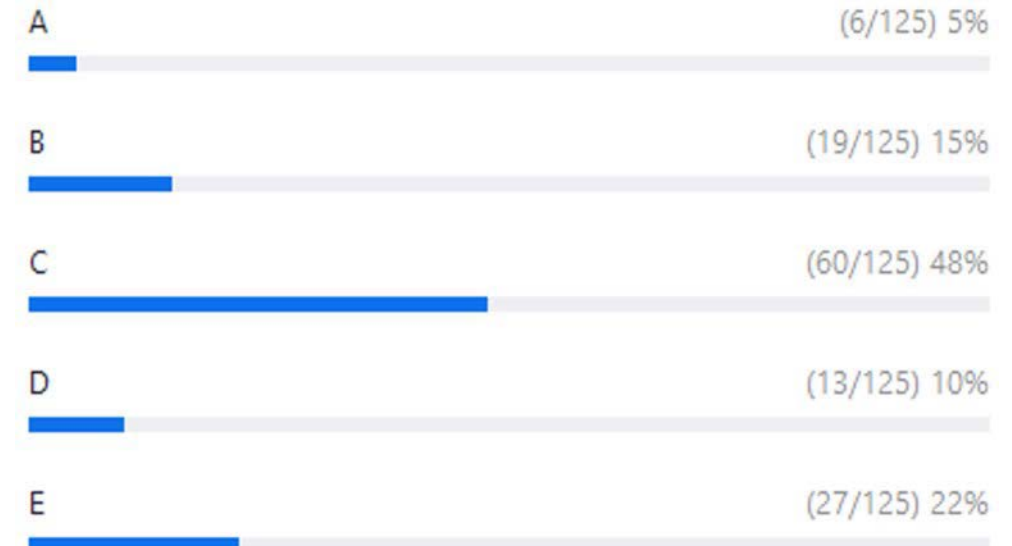
Updated 4:36 PM ET, Thu February 28, 2019



Based on the news headline, which of the following is true?

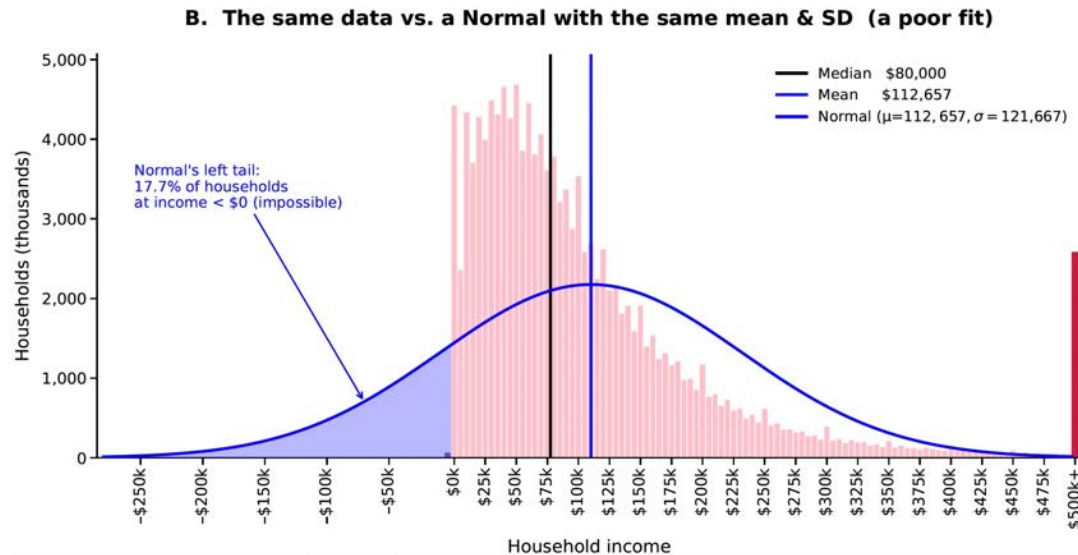
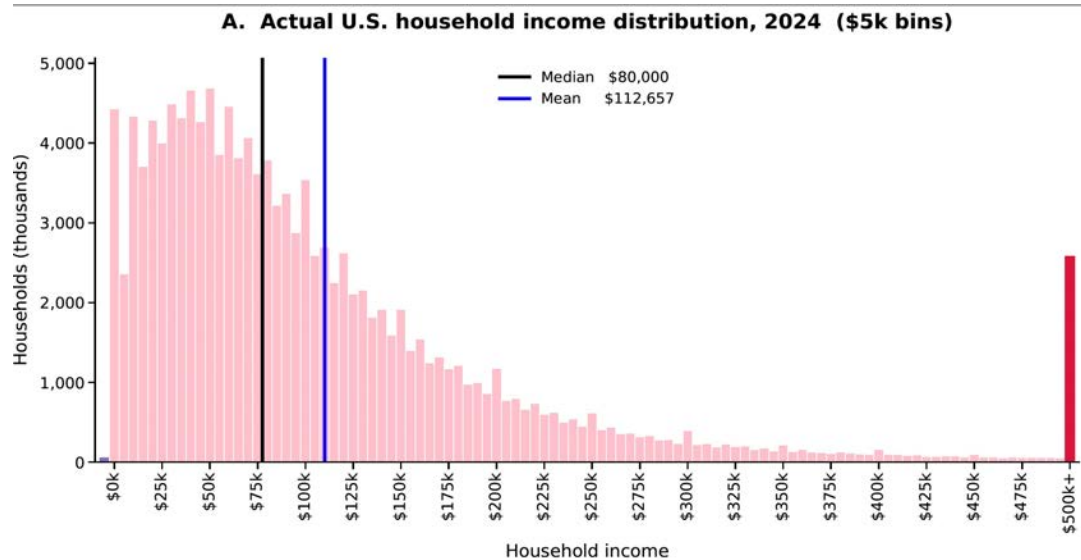
- A. I can expect around \$40 more this year
- B. I can expect to get a bigger refund, amount uncertain
- C. Many people will get approximately \$40 more, but “individual results may vary”
- D. A typical refund will be \$40 more
- E. None of the above

class poll:



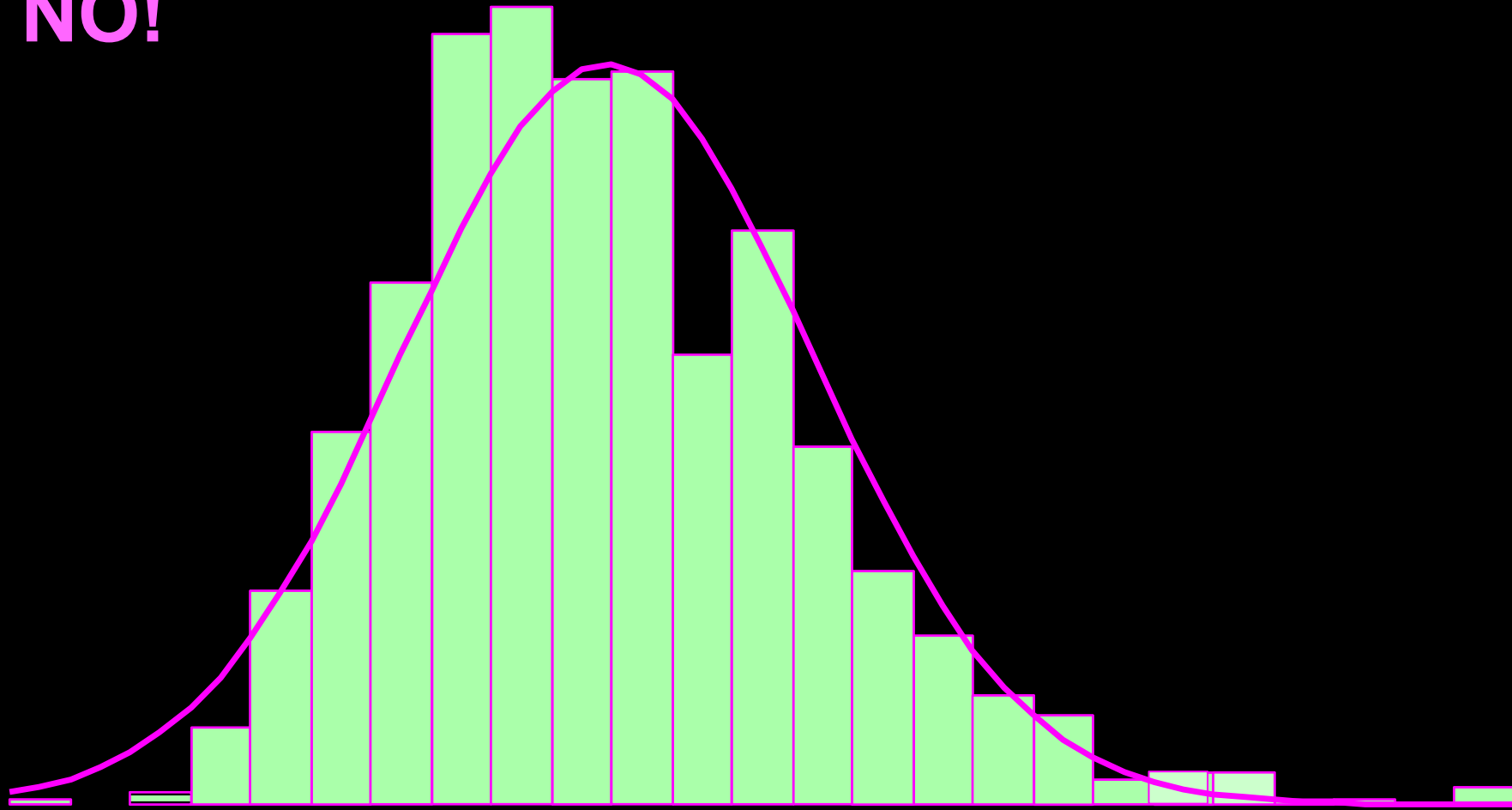
Why did so many people get that wrong?

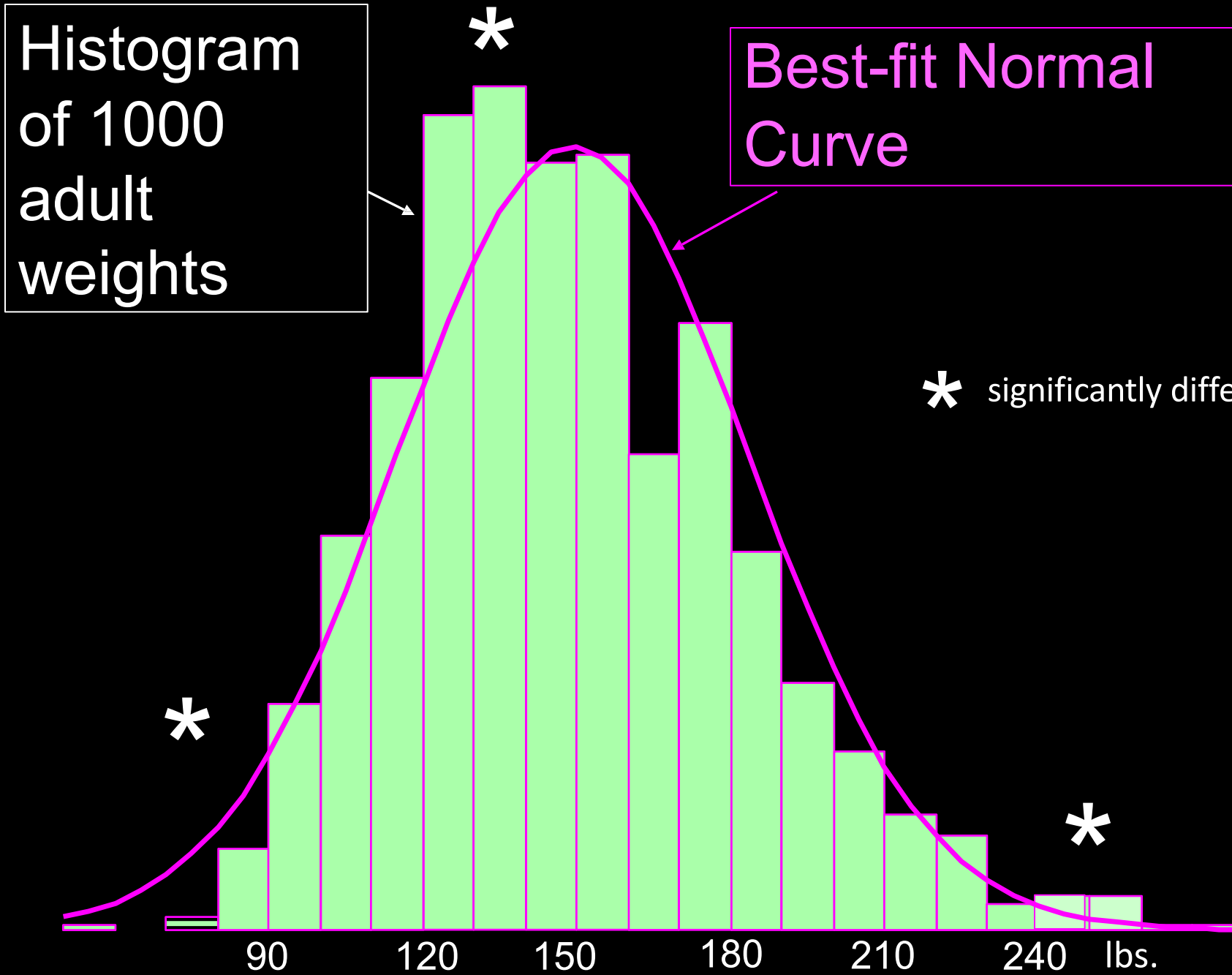
A: Because they unconsciously assumed a Normal distribution

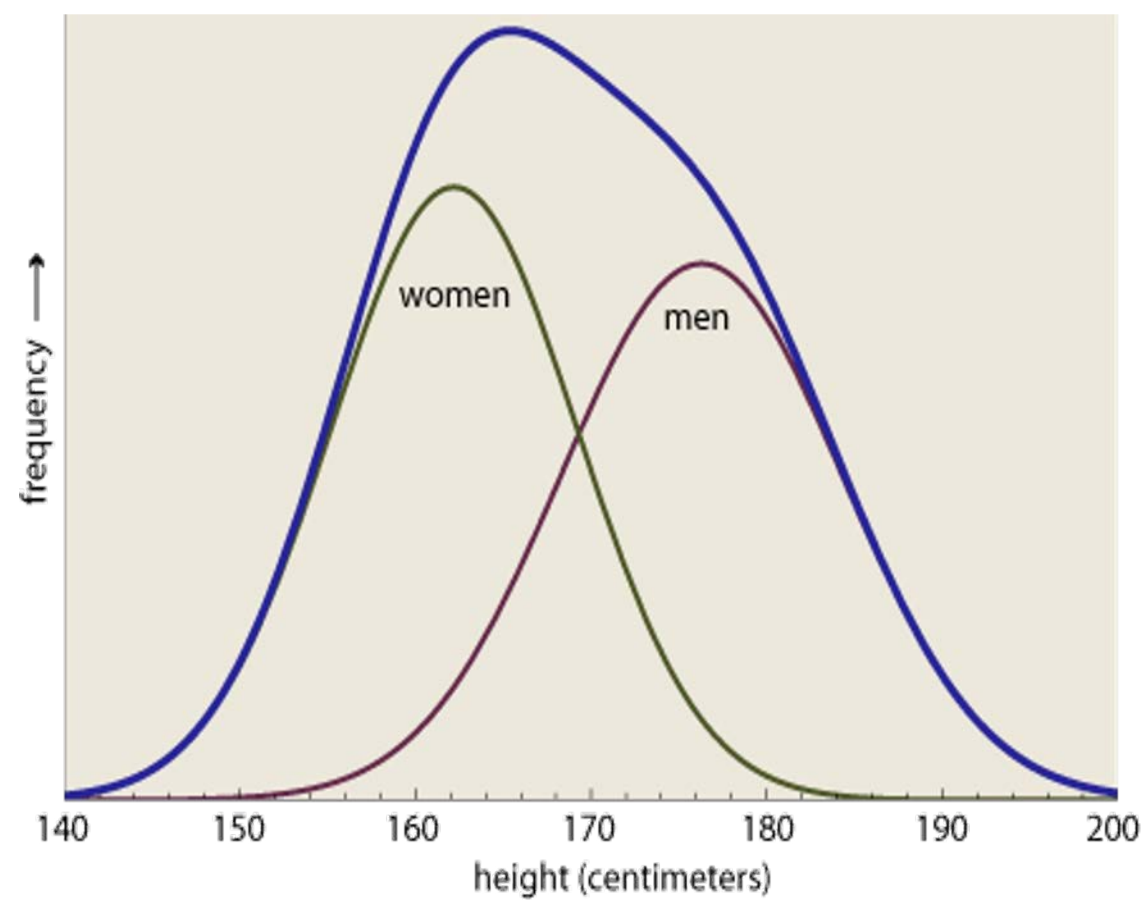


are real world distributions normal?

NO!



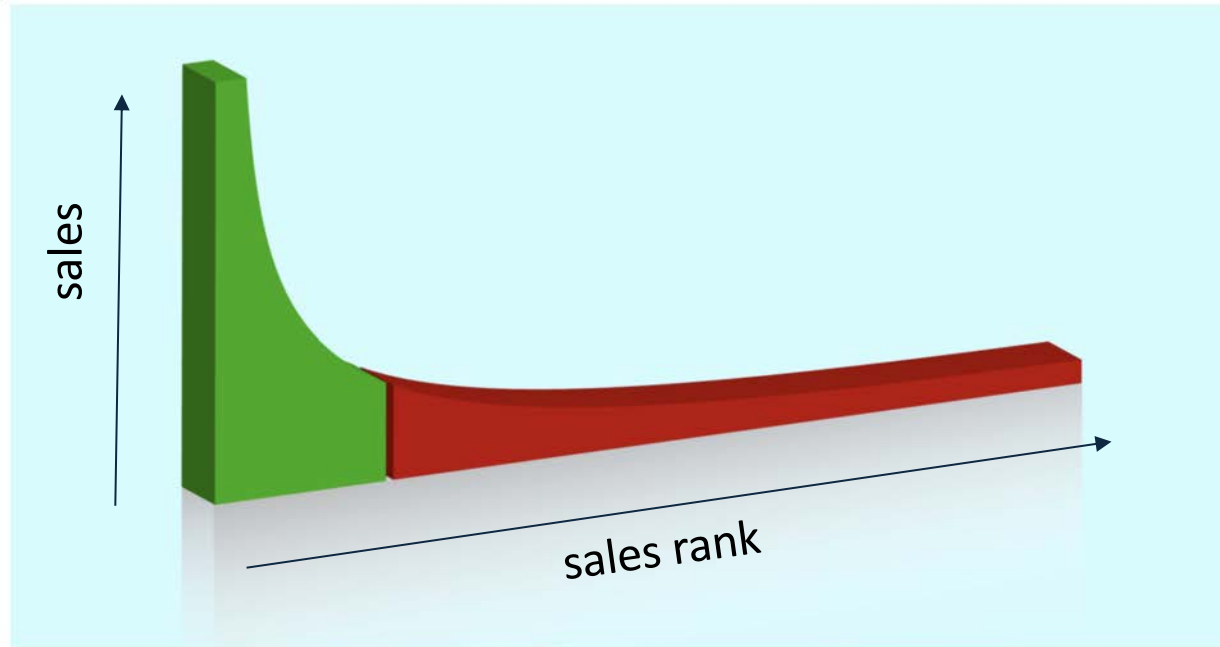




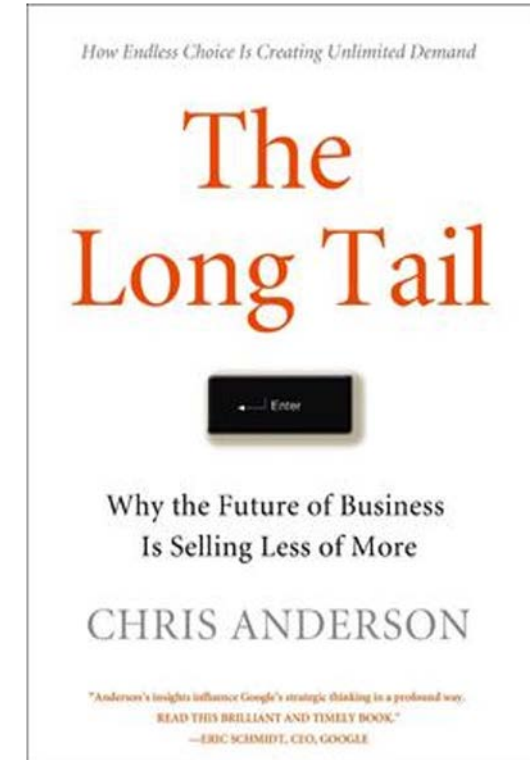
Why aren't “real” quantities “normal”?

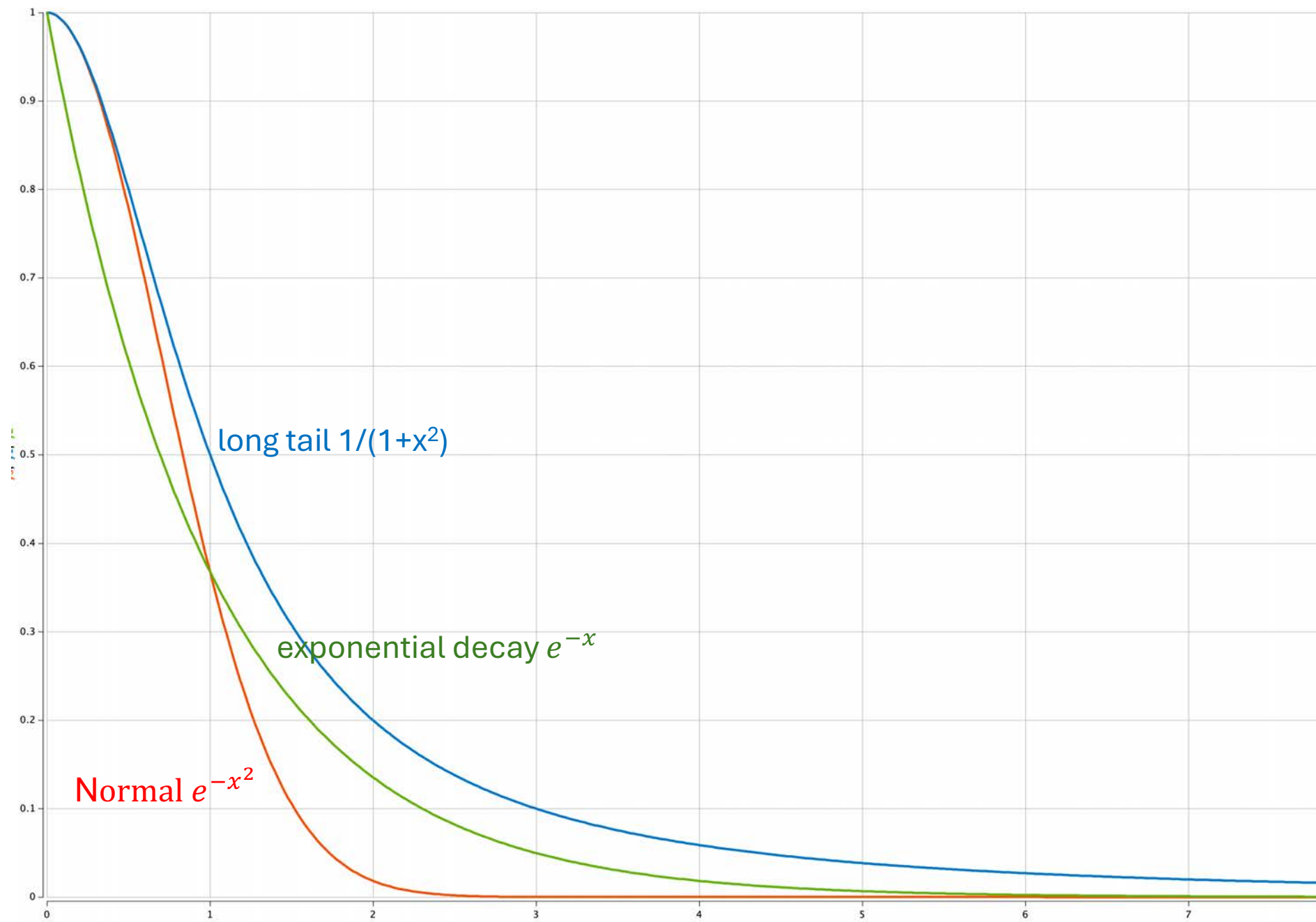
1. They are “mixtures” of distributions, for example, combining males and females
2. Most biological quantities are “bounded” (e.g., they cannot be negative)
3. Most distributions have naturally “long” or “thick” tails

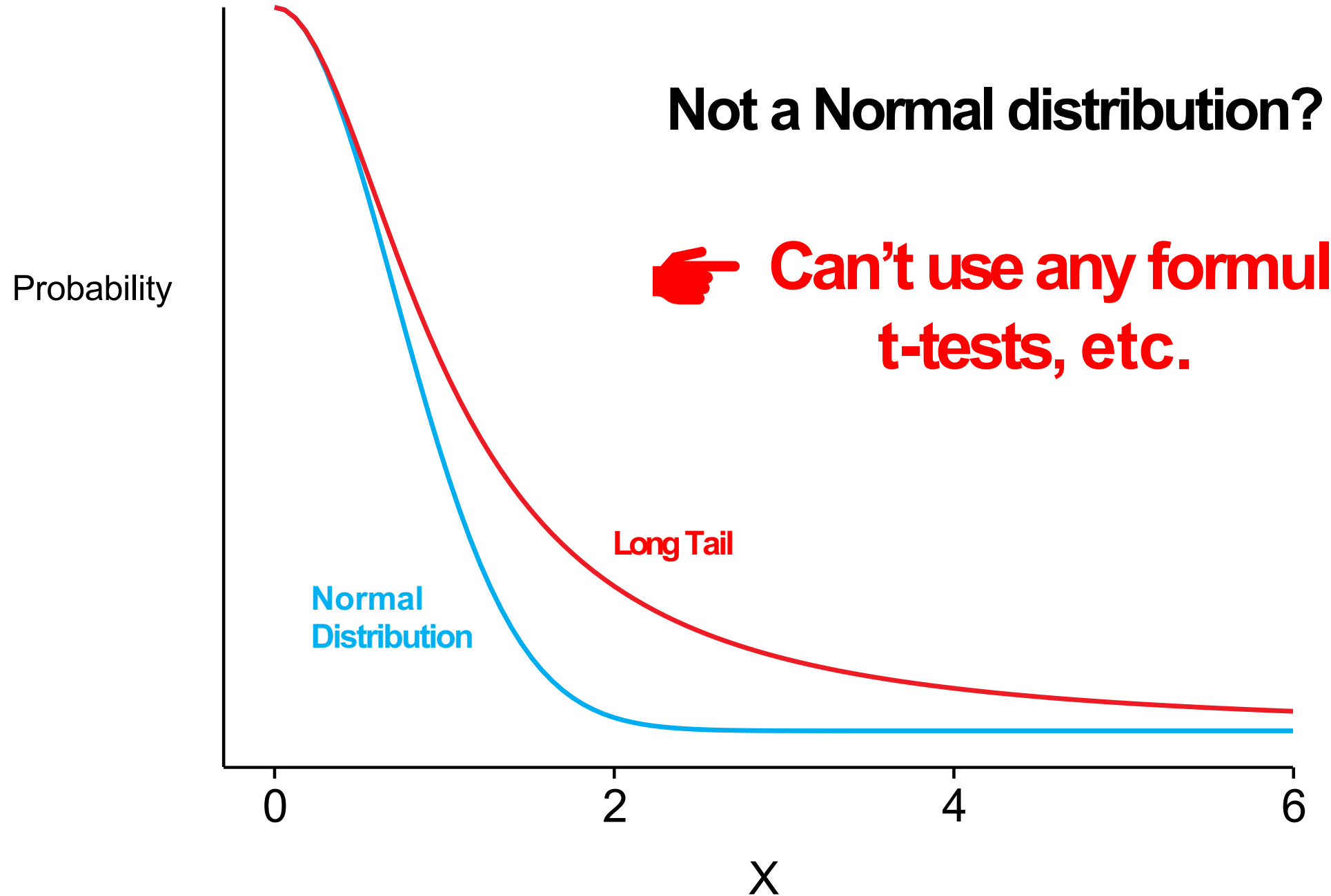
Thick tails and long tails



- books at Amazon.com
- songs at iTunes
- followers of influencers on social media







THE FLAW OF AVERAGES

Todd Rose “The Flaw of Averages” (Toronto Star)

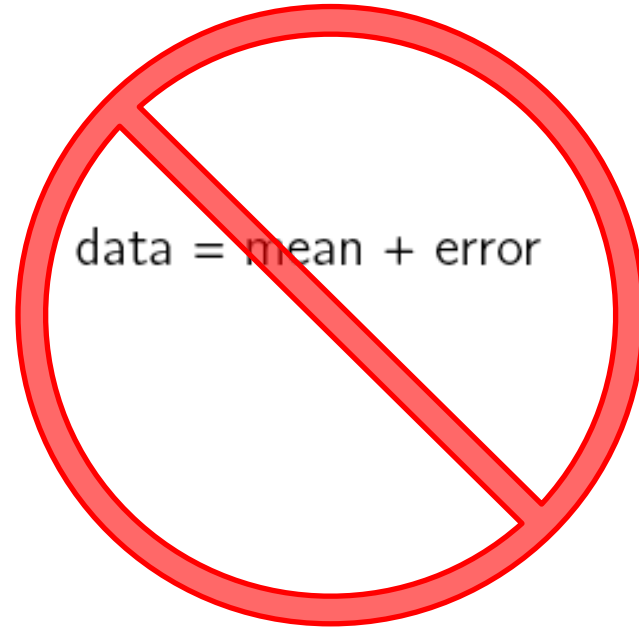
"Norma": the perfect woman? In Rose's account, the obsession with the average became a kind of national psychosis. "A doctor in Cleveland took the averages of dozens of body measurements on thousands of women, and proclaimed that the averages represented the ideal woman. They even had a sculptor make a statue, called "Norma", based on these average measurements. . . . Norma was featured in Time magazine, in newspaper cartoons, and on an episode of a CBS documentary series, This American Look, where her dimensions were read aloud so the audience could find out if they, too, had a normal body." (Figure 2.37).

the doctor in question was Dr. Robert L. Dickinson, president of the American Gynecological Society and chairman of obstetrics at the American Medical Association.



Norma, based on average measurements of size data collected from 15,000 young adult women.

The psychosis of “the average” lent itself to what is probably the worst statistical idea of the 20th century: the idea that population variability is a kind of “noise” or “error”, and that we can use the average value as representative of the distribution.



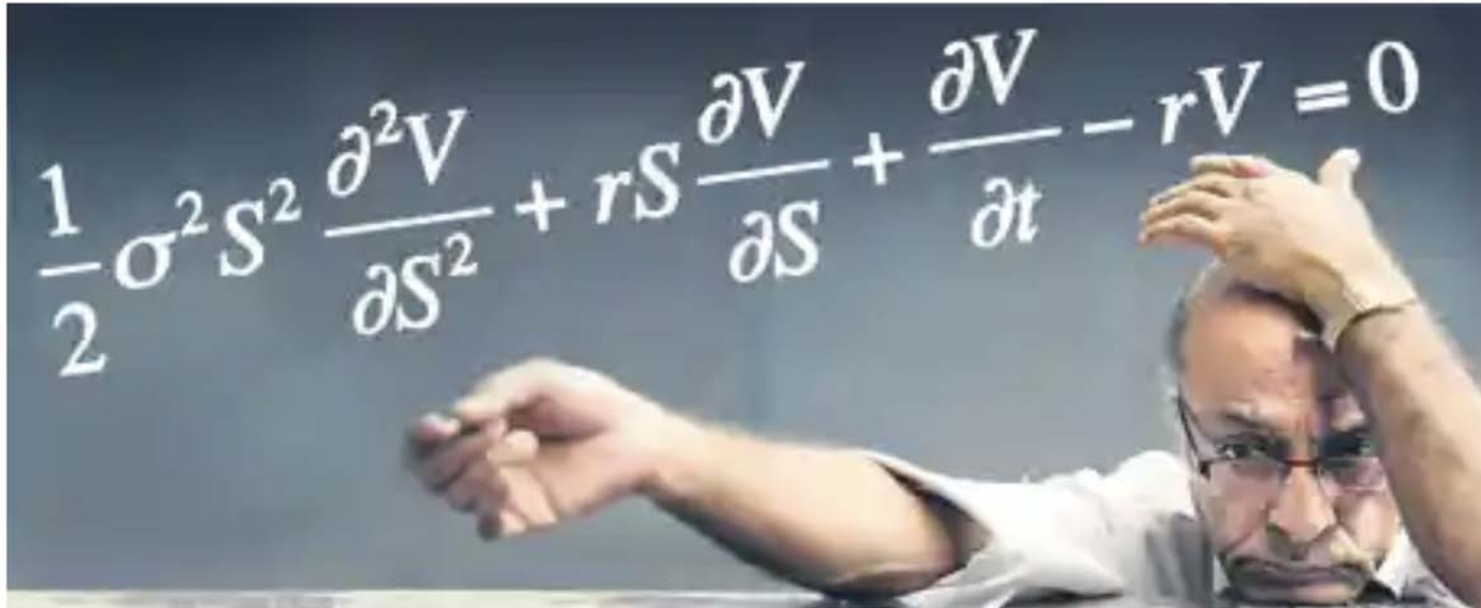
Treating the average patient The flaw of averages, the idea that a population can be characterized by an average value plus “error”, has disastrous consequences for clinical treatment. A telling demonstration of this fact was shown by Kent and Hayward (2007), in their article “When Averages Hide Individual Differences in Clinical Trials”. They showed the work of artist Chris Dorley-Brown, who took 2,000 photographs of people living in a small town in England, and created a digital “average face” (center photograph in Figure 2.38). The average person is neither male nor female, neither young nor old, and neither sick nor healthy. Medical therapy tailored to the average person will miss many important clinical dimensions.



Figure 2.38 Reprint of Figure 1 from the article “When Averages Hide Individual Differences in Clinical Trials” by Kent and Hayward (2007).

The mathematical equation that caused the banks to crash

The Black-Scholes equation was the mathematical justification for the trading that plunged the world's banks into catastrophe

A photograph of a man with glasses and a white shirt, pointing his right hand towards the Black-Scholes equation written on a chalkboard. The equation is written in white chalk on a dark background. The man's left hand is raised near his head. The equation is:
$$\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} - rV = 0$$

📷 In the Black-Scholes equation, the symbols represent these variables: σ = volatility of returns of the underlying asset/commodity; S = its spot (current) price; δ = rate of change; V = price of financial derivative; r = risk-free interest rate; t = time. Photograph: Asif Hassan/AFP/Getty Images



Ian Stewart

The Observer, Saturday 11 February 2012

Black-Scholes Equation


$$\frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} - rV = 0$$

V = price of option

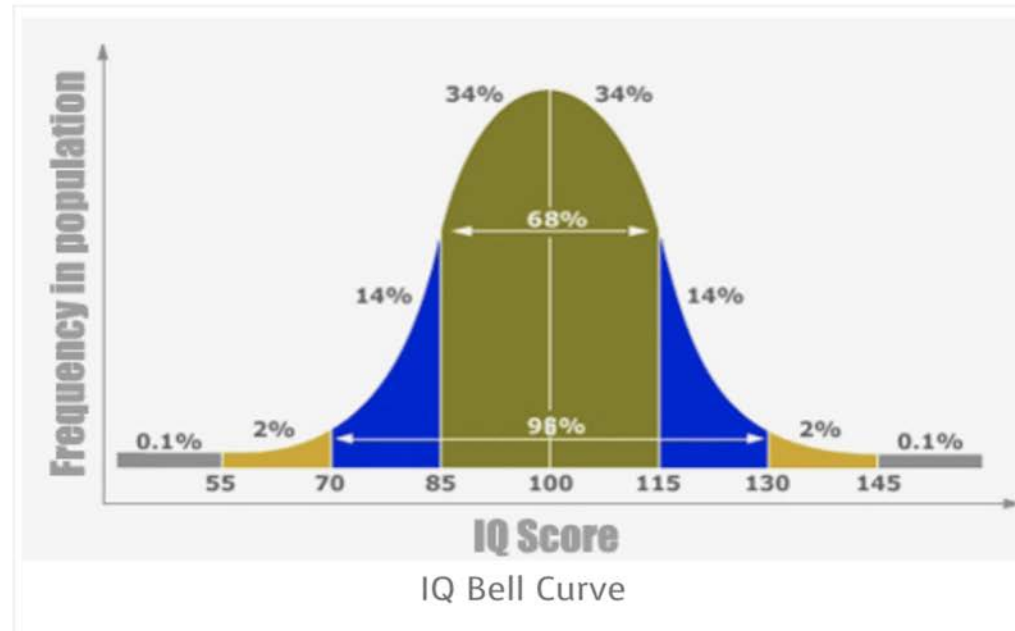
S = stock price

r = interest rate

σ = “volatility” (= standard deviation of price fluctuations)

The IQ Score Bell Curve – A 'Normal Distribution'

Standardized IQ tests are designed so that their scores have a 'bell curve' distribution in the general population with an average of 100. This curve has a peak in the middle where most people score and tapering ends where only a few people score. In statistics this is called a ***normal distribution***.



“The IQ scale is *defined* so that it gives a normal distribution with mean 100 and standard deviation 15. **This is possible for any test score with a continuous distribution f .** If the subject's score in the test is s , their IQ will be given by:

$$\text{IQ} = 100 + 15\sqrt{2} \operatorname{erfc}^{-1} \left(2 - 2 \int_{-\infty}^s f(x) dx \right)$$

“You can always map the scores so that any distribution, whatever it is, becomes normal.”
“IQ is normal by definition”

To illustrate how you'd calculate IQ in practice, I made a Python script which takes an arbitrary distribution of scores in a test, generates a sample of 10 thousand results and uses that to calculate the IQ of an additional participant based on their score in the same test. The plot shows the distribution of scores and how it's transformed into a normal distribution when converted to IQ. The score and IQ of the new participant is shown in red.

```
import bisect
import numpy as np
from scipy.special import erfcinv
from scipy.stats import rv_continuous

N_SAMPLES = 10000

np.random.seed(0)

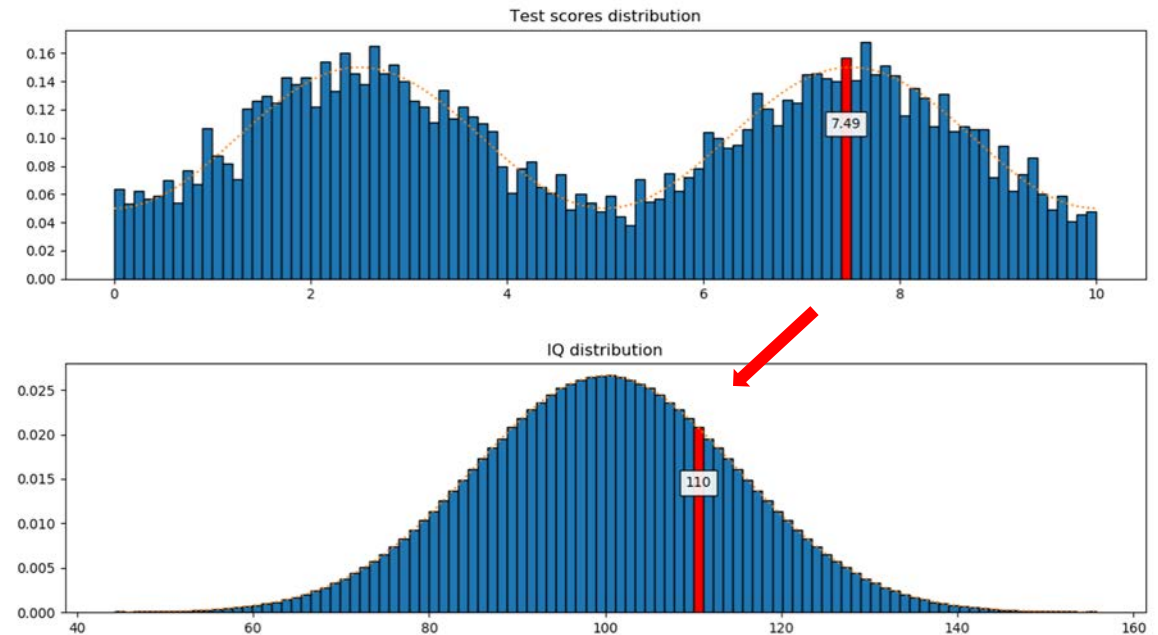
# Start with any continuous distribution for the test scores.
# In this case, it's a multimodal distribution.
def scores_pdf(x):
    return (2-np.cos(2*np.pi*x/5))/20

class dist_gen(rv_continuous):
    def _pdf(self, x):
        return scores_pdf(x)

# We restrict the score to be between zero and ten.
dist = dist_gen(a=0, b=10)

scores = dist.rvs(size=N_SAMPLES)
scores = list(sorted(scores))

# Convert from percentile to IQ
```

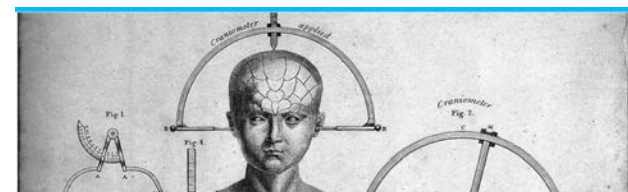
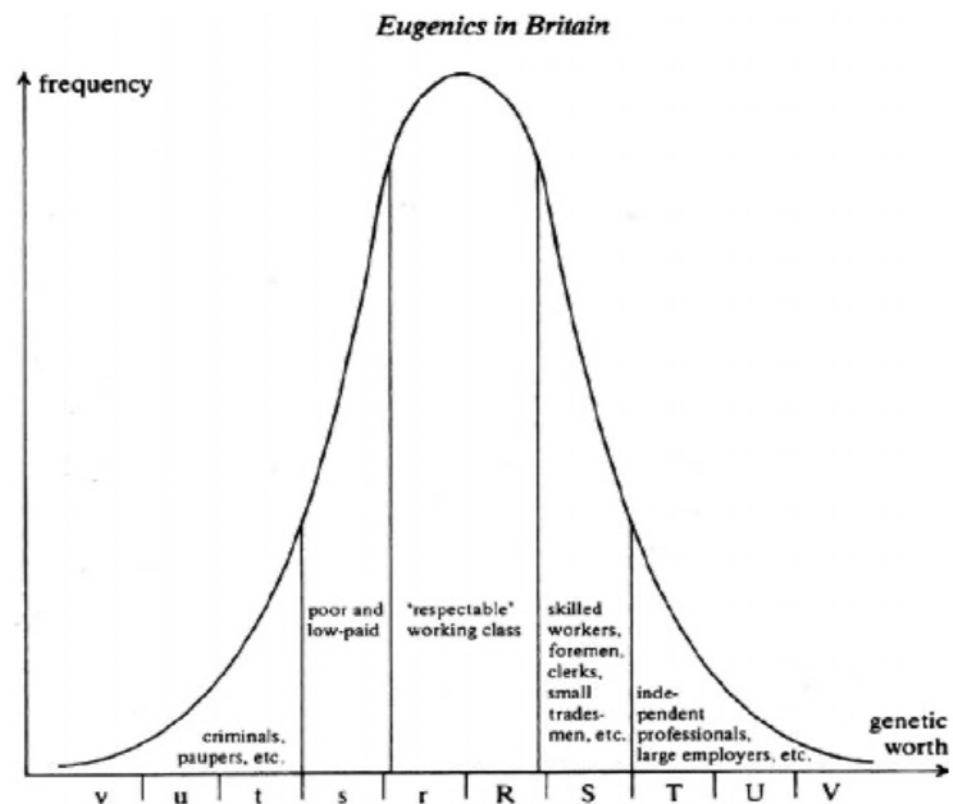


<https://math.stackexchange.com/questions/1311149/why-are-iq-test-results-normally-distributed>

Therefore, the assertion that “IQs are Normally distributed” is a phony “discovery”: the Normal Distribution is an artifact of the test construction.

Why did people ever think that there was something called “intelligence” that was distributed Normally? The eugenicist Francis Galton simply pontificated it in a wild and false extrapolation. He said (1869):

“Now, if this [normal distribution] be the case with stature, then it will be true as regards every other physical feature — circumference of head, size of brain, weight of grey matter, number of brain fibres, &c.; and thence, by a step on which no physiologist will hesitate, as regards mental capacity.”



How Eugenics Shaped Statistics

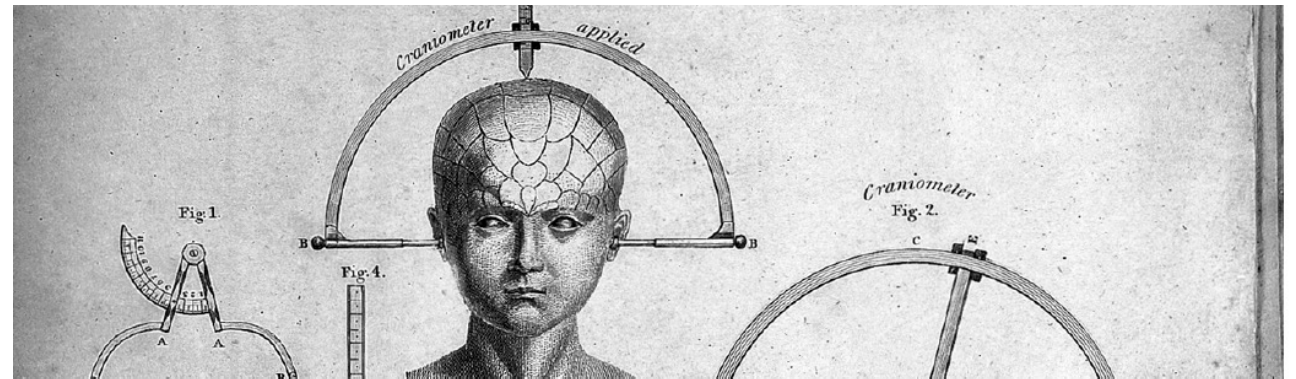
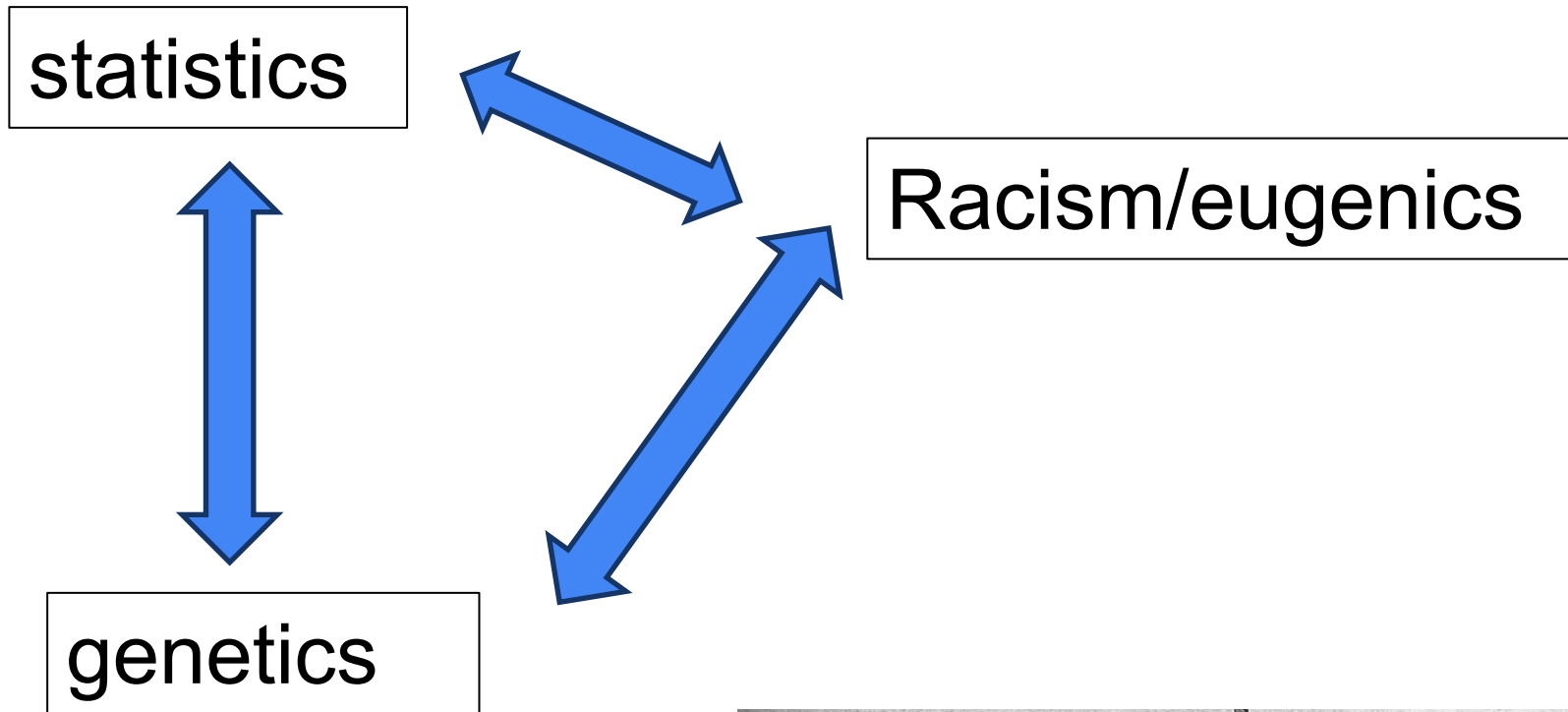
Exposing the damned lies of three science pioneers.

BY AUBREY CLAYTON
OCTOBER 14, 2020

BERNOULLI'S
FALLACY

Statistical Illogic
and the Crisis of
Modern Science

AUBREY CLAYTON



How Eugenics Shaped Statistics

Exposing the damned lies of three science pioneers.

BY AUBREY CLAYTON
OCTOBER 28, 2020

Take-home lessons

- **Distributions in nature are NOT “Normal”**
- **Assumptions based on distributions being Normal are dangerous and wrong**
- **Techniques that were derived for Normal distributions (means, standard deviations, “t-tests” etc.) are in error**

19th – mid 20th Century

- formulas
- paper-and-pencil “solutions”
- Normal distributions only
- means and standard deviations

***t*-test**

$$t = \sqrt{n} \left(\frac{\bar{X}_A - \bar{X}_B}{\hat{\sigma}_{AB}} \right)$$

$$\rho(t | \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu \pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu} \right)^{-\frac{\nu+1}{2}}$$

$$\nu = n - 1$$

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$$

***χ*²-test**

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

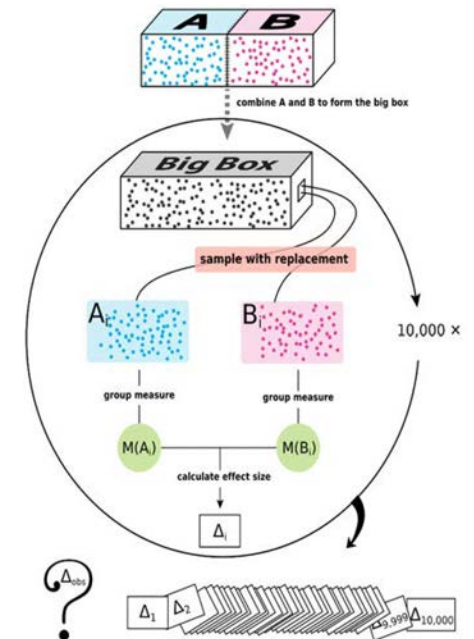
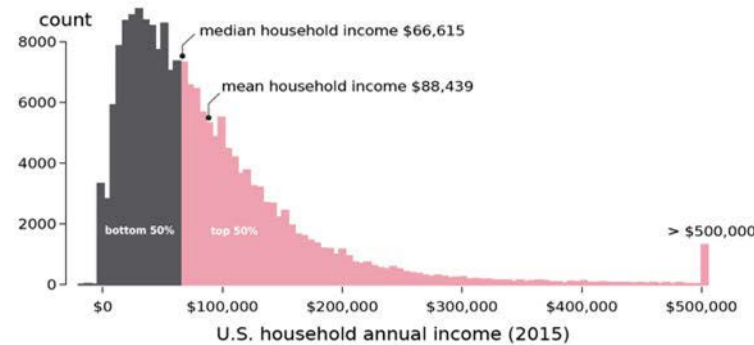
$$\chi^2 = \frac{1}{2^{d/2} \Gamma(d/2)} x^{d/2-1} e^{-x/2}$$

where $\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$

Statistics

late 20th Century

- shapes of distributions
- all distributions welcome
- measures other than the mean and standard deviation
- computer simulation



distributions not normal → no formulas

Statistical Data Analysis in the Computer Age

BRADLEY EFRON AND ROBERT TIBSHIRANI Stanford Statistics

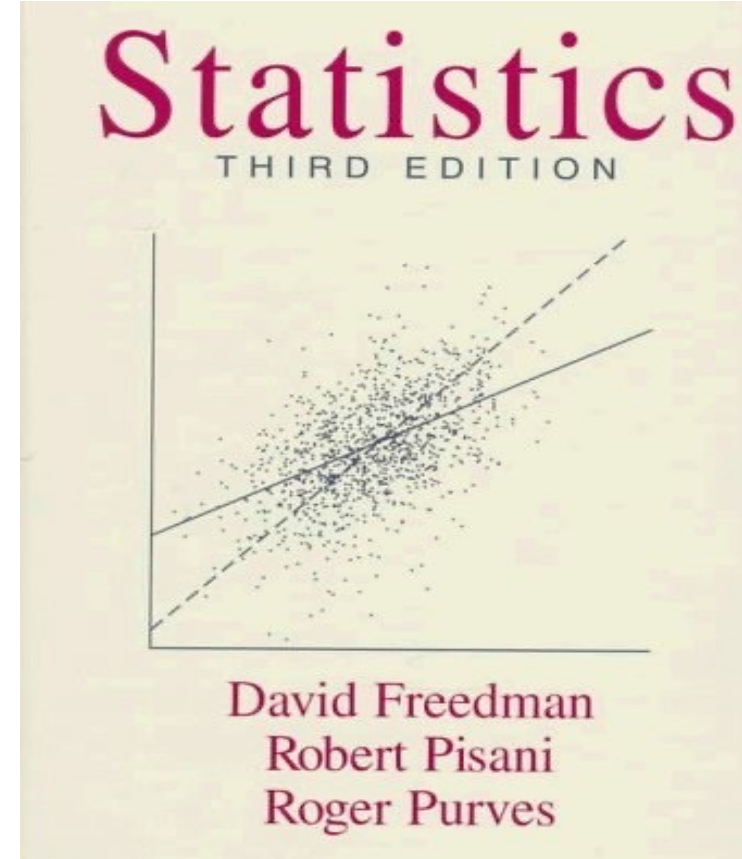
“Most of our familiar statistical methods...were designed to be implemented on mechanical calculators. Modern electronic computation has encouraged a host of new statistical methods that **require fewer distributional assumptions** than their predecessors and can be **applied to more complicated statistical estimators.**”

non-Normal



not just
the mean

“In Pearson’s time, there were no computers to find the chances. So he developed a method for approximating P by hand ... without having to do a computation that was, — by the standard of his time, quite formidable.”



Freedman D, Pisani R, Purves R. (UC Berkeley)
2007



Comparing Two Groups

A	B
a_1	b_1
a_2	b_2
$\vdots$	$\vdots$
a_{49}	b_{45}

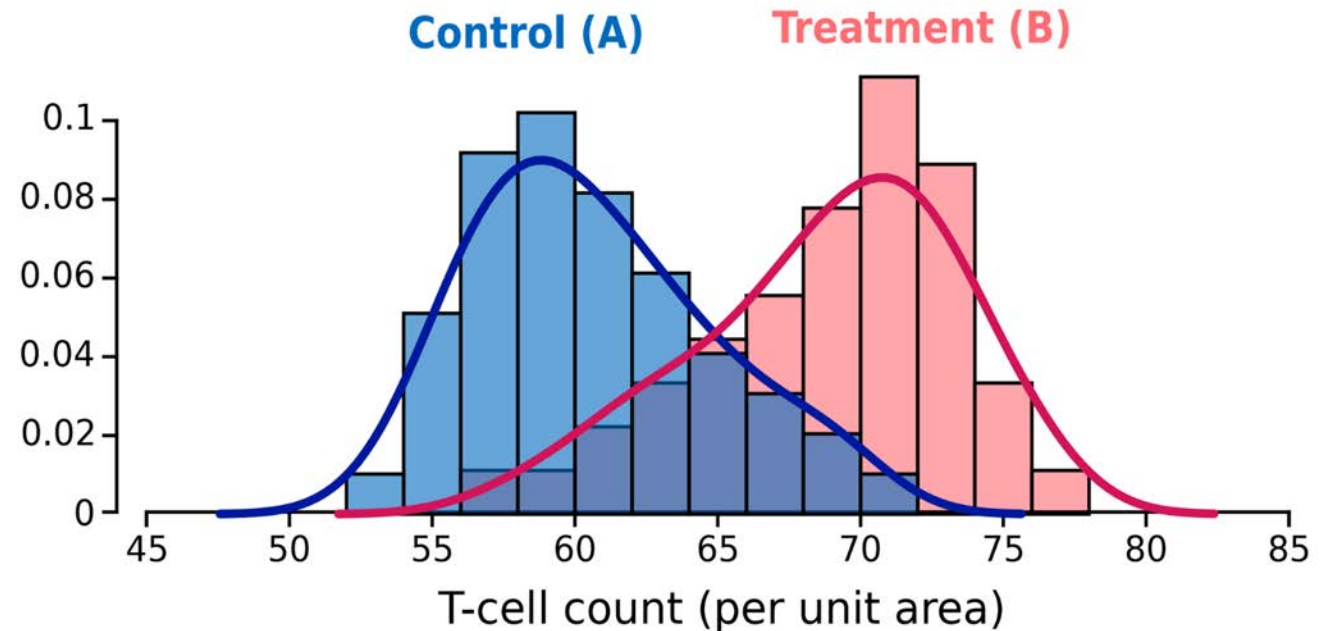
note $n_A \neq n_B$

T-cell activation by Drug

T-cells cultured in a dish. Outcome measure: T-cell count/unit area

Step #1: Visualize the Data

Control group	Treatment group
A	B
a_1	b_1
a_2	b_2
$\vdots$	$\vdots$
a_{49}	b_{45}



- not Normally distributed
- is the average (mean) value the right description of this data?

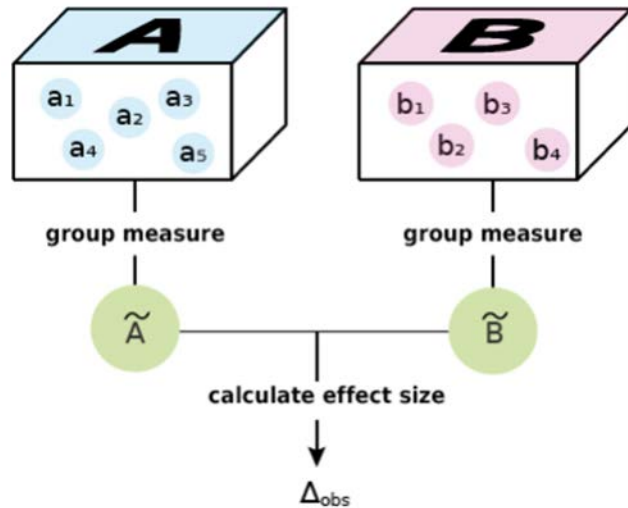
The median gives you actual information, the mean (average) often does not.

You are a patient, told: “The **average** survival time with your type of cancer is 2 years.”

You are a patient, told: “The **median** survival time with your type of cancer is 2 years.”

Because “median” means “half the people”, it implies that you have a 50/50 chance of making it past 2 years

Step #2: choose appropriate group measure M, and choose method of comparison (subtraction? ratio?) to get effect size



Effect size

$$\Delta_{obs} = M(B) - M(A) = \tilde{B} - \tilde{A}$$

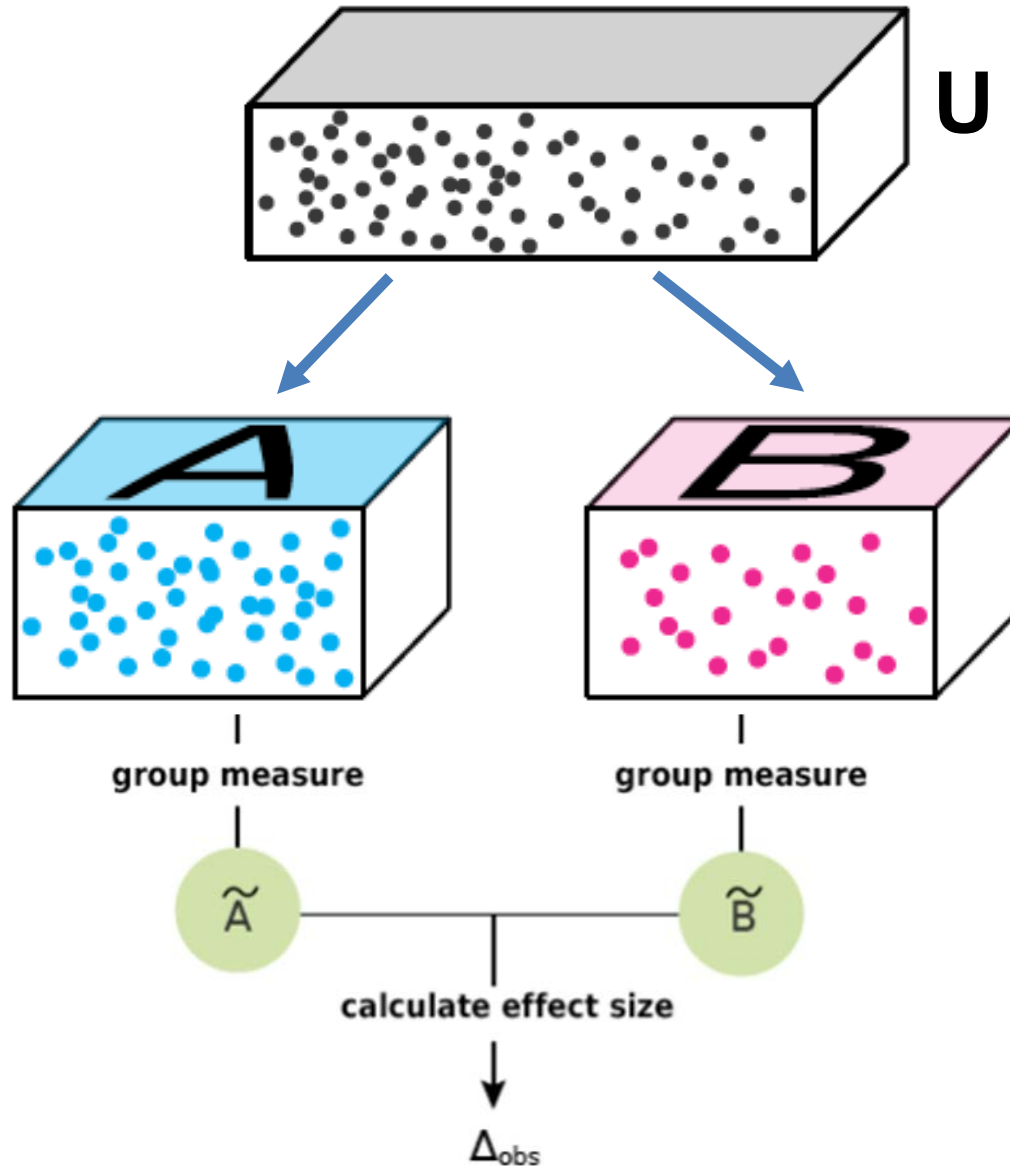
Is Δ_{obs} “large”???

large relative to what?

relative to what we might expect from pure randomness

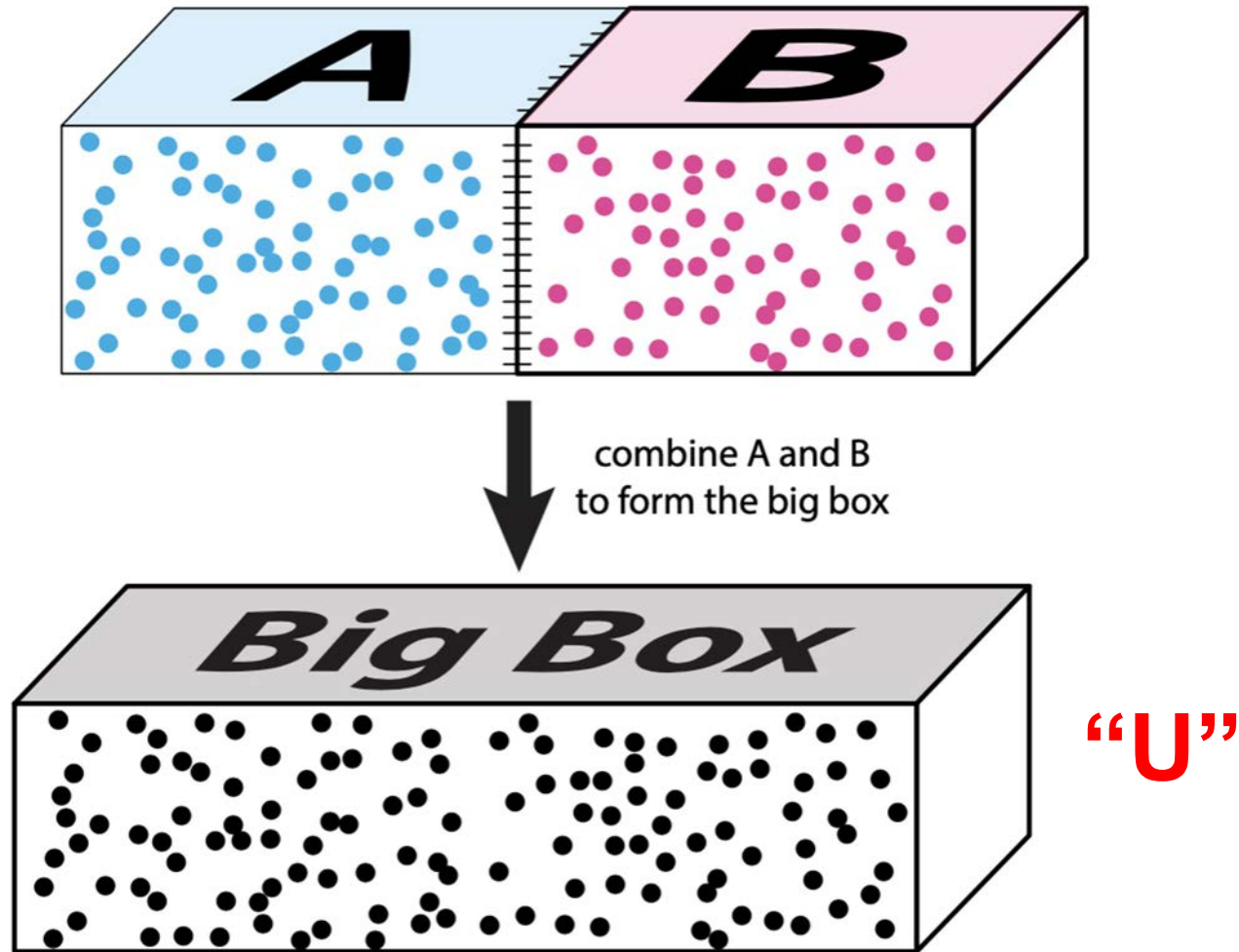
**How big a difference would you expect to see if
A and B were just 2 random samples from the
same set?**

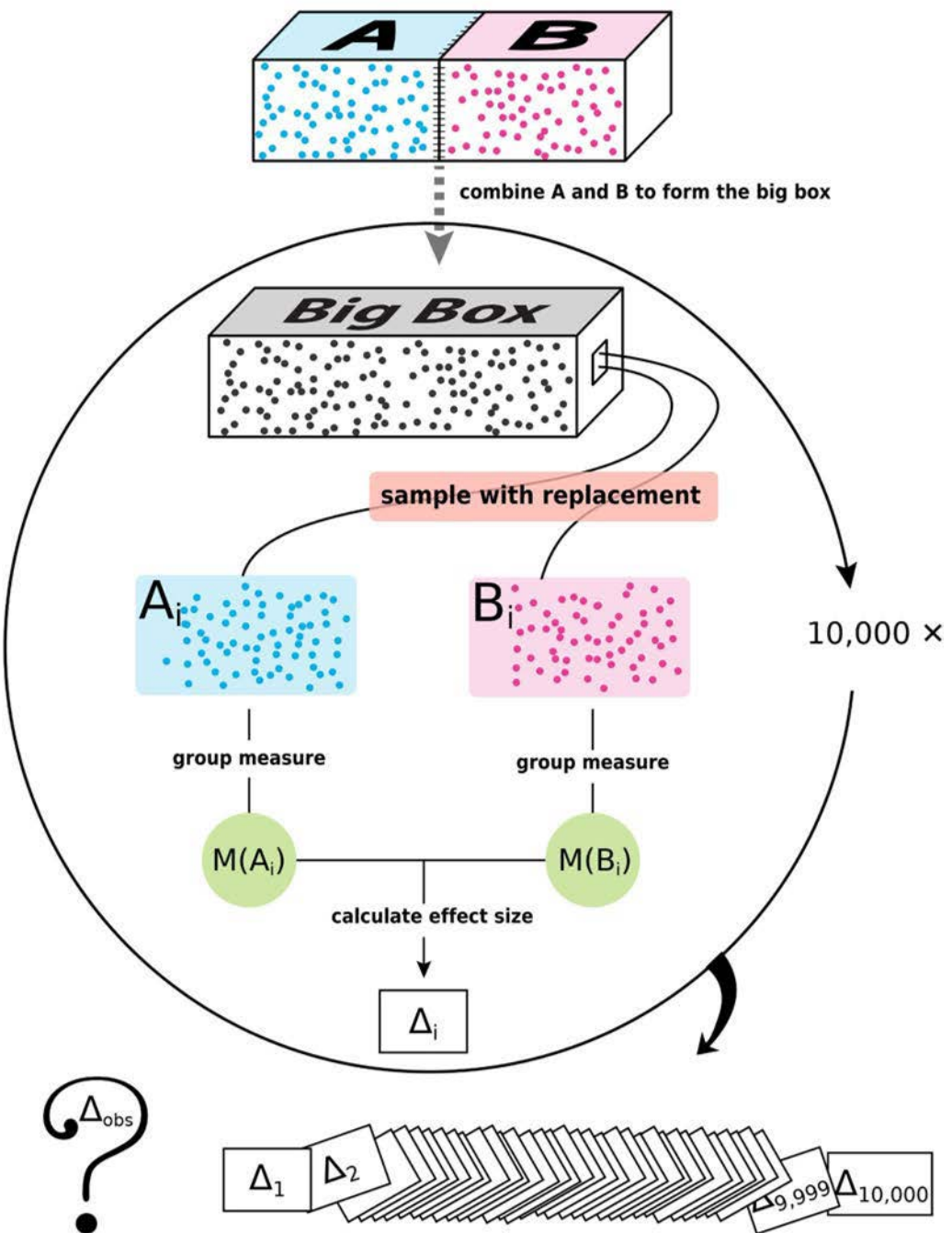
couldn't A and B just be 2 samples drawn from the same distribution U??



Any pair of samples drawn at random will have some difference!

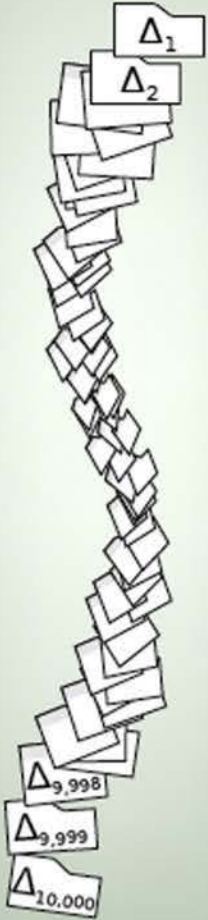
In fact, the only thing we can say about reality, about the underlying distribution U , is: *it must look like our data, since our data are drawn from it*. We are, as always, assuming that our samples A and B are not biased, and therefore, the best guess about what U looks like is to put A and B into a big box: that's our best model of U .





Resampling Procedure

10,000 differences
from resamples



sort



10,000 differences from resamples (ranked)

largest

Rank

1
2
⋮
49
50
51
⋮
4,999
5,000
5,001
⋮
9,950
9,951
9,952
⋮
9,999
10,000

Δ_{121}
 $\Delta_{8,052}$
⋮
 Δ_7
 Δ_{91}
 Δ_{143}
⋮
 Δ_{74}
 $\Delta_{9,910}$
 Δ_1
⋮
 $\Delta_{1,945}$
 Δ_{618}
 $\Delta_{7,374}$
⋮
 Δ_{12}
 $\Delta_{1,560}$

what is the
p-value ?

?

?

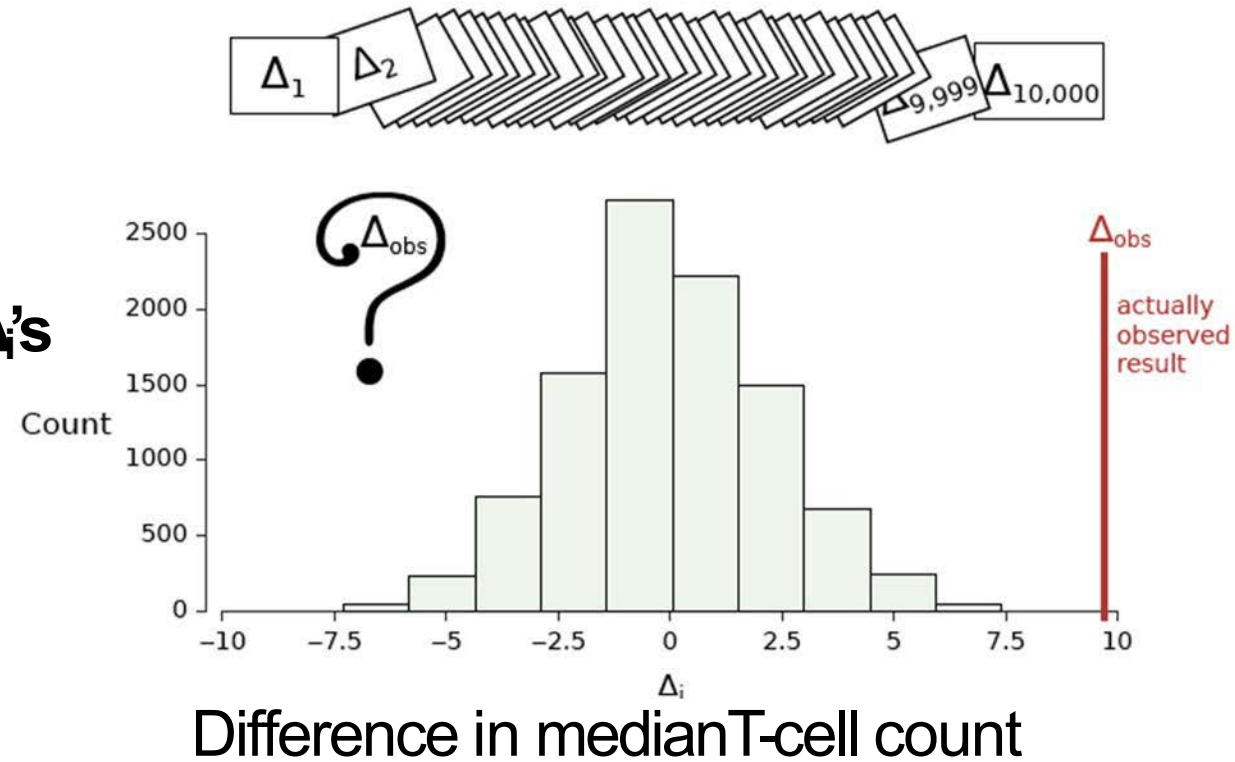
?

Δ_{obs}
(observed difference)

Where is your Δ_{obs}
in the ranked list of
10,000 Δ_i values
from the
resamplings?

Count number of
simulations as or more
extreme than Δ_{obs} to
calculate p-value.

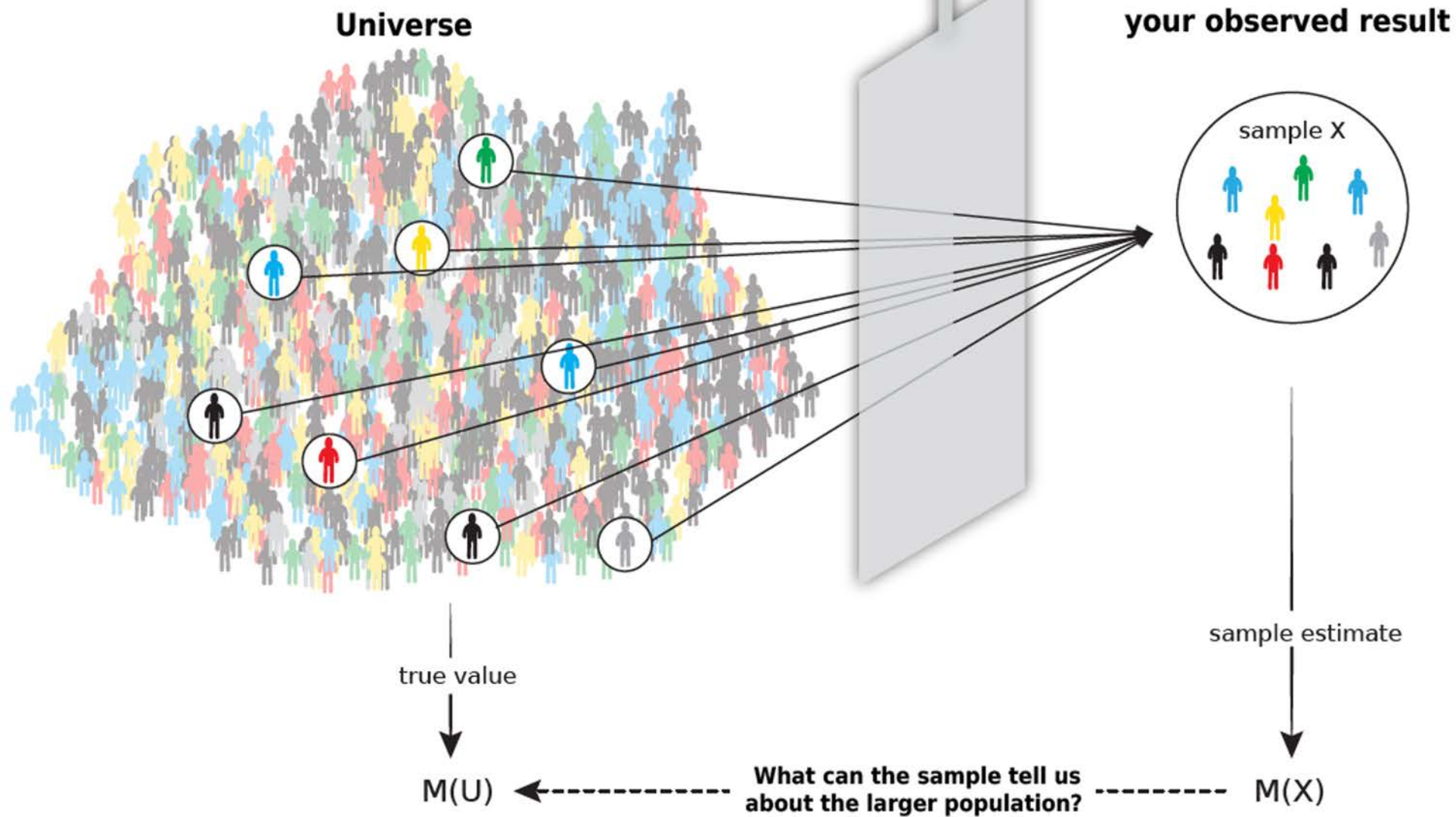
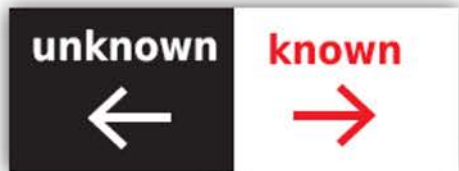
Histogram of Δ 's



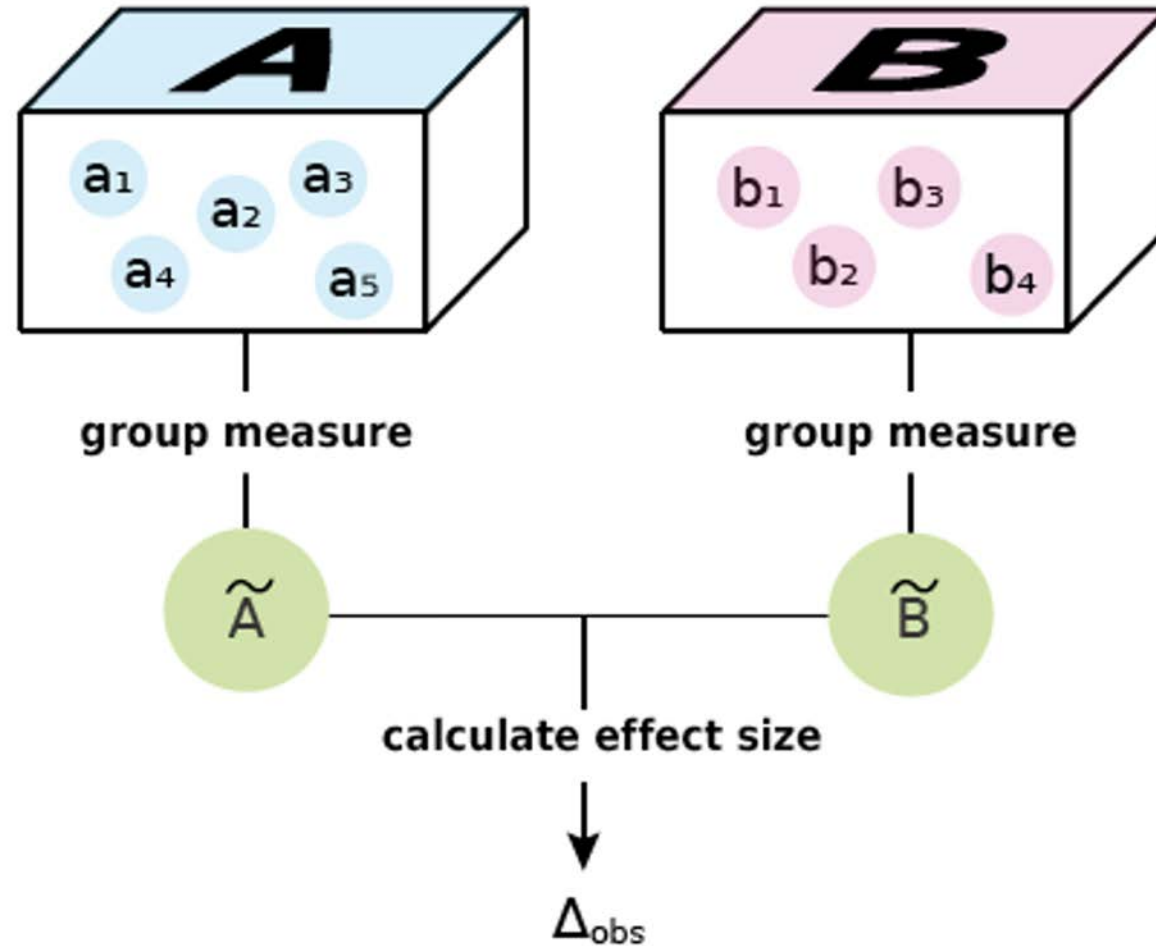
Result: The median T-cell count was improved under the drug by 9.7 cells per unit area ($p < 0.0001$).

Confidence Intervals for the effect size: much better!

Better than just p-values + effect size

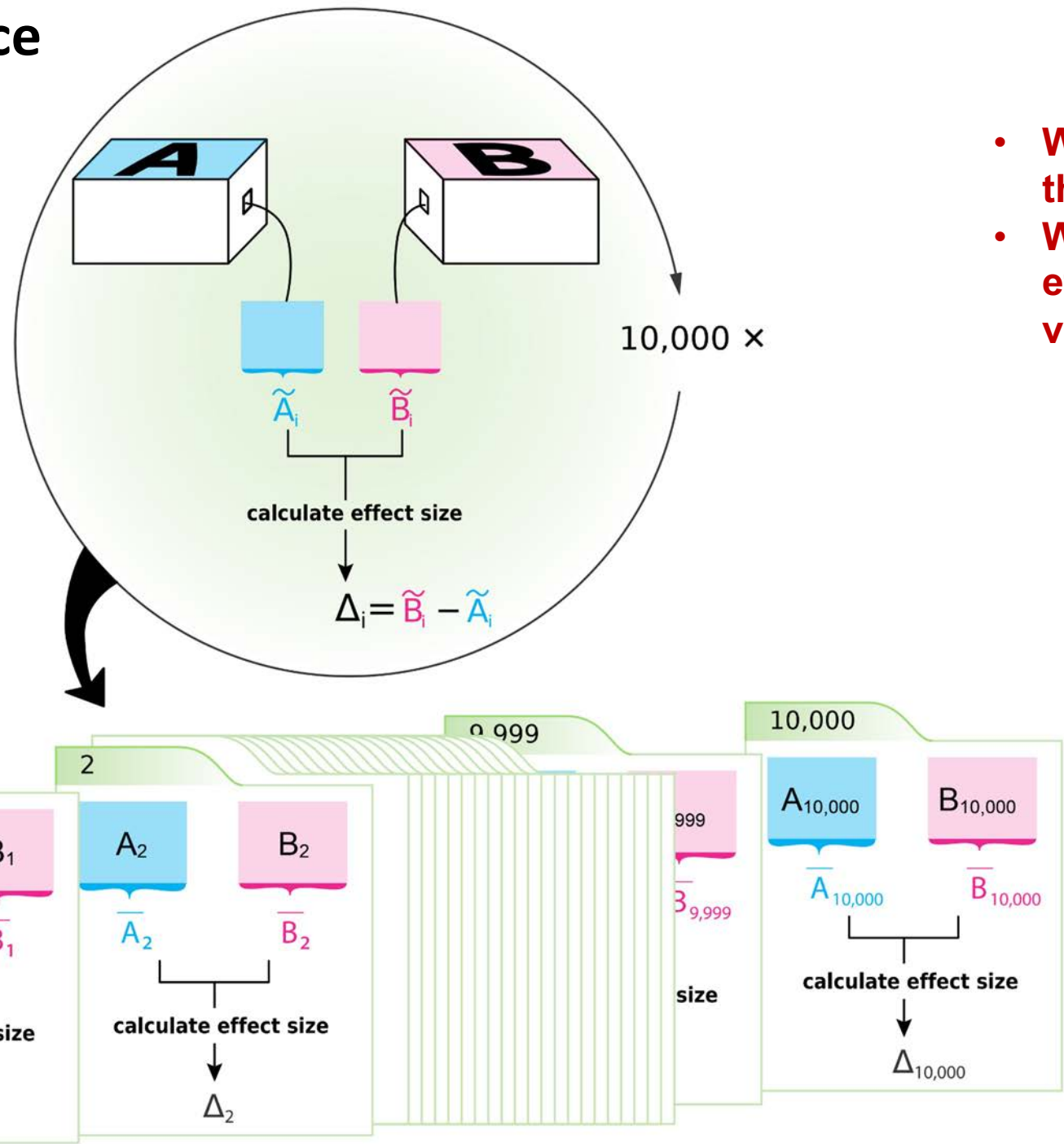


Confidence Intervals- This is what we really want: get a confidence interval for the effect size



Construct Confidence Interval Interval for Effect Size

Box model =
“Two-Box”

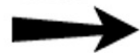


- We are NOT simulating the Null Hypothesis
- We are simulating the effect of sampling variability on Δ_{obs}

10,000 differences
from resamples



sort



smallest

largest

10,000 differences from resamples (ranked)

Rank

1
2
⋮
49
50
51
⋮
4,999
5,000
5,001
⋮
9,950
9,951
9,952
⋮
9,999
10,000

Δ_{121}

$\Delta_{8,052}$

⋮

Δ_7

Δ_{91}

Δ_{143}

⋮

Δ_{74}

$\Delta_{9,910}$

Δ_1

⋮

$\Delta_{1,945}$

Δ_{618}

$\Delta_{7,374}$

⋮

Δ_{12}

$\Delta_{1,560}$

bottom
0.5%

Δ_L

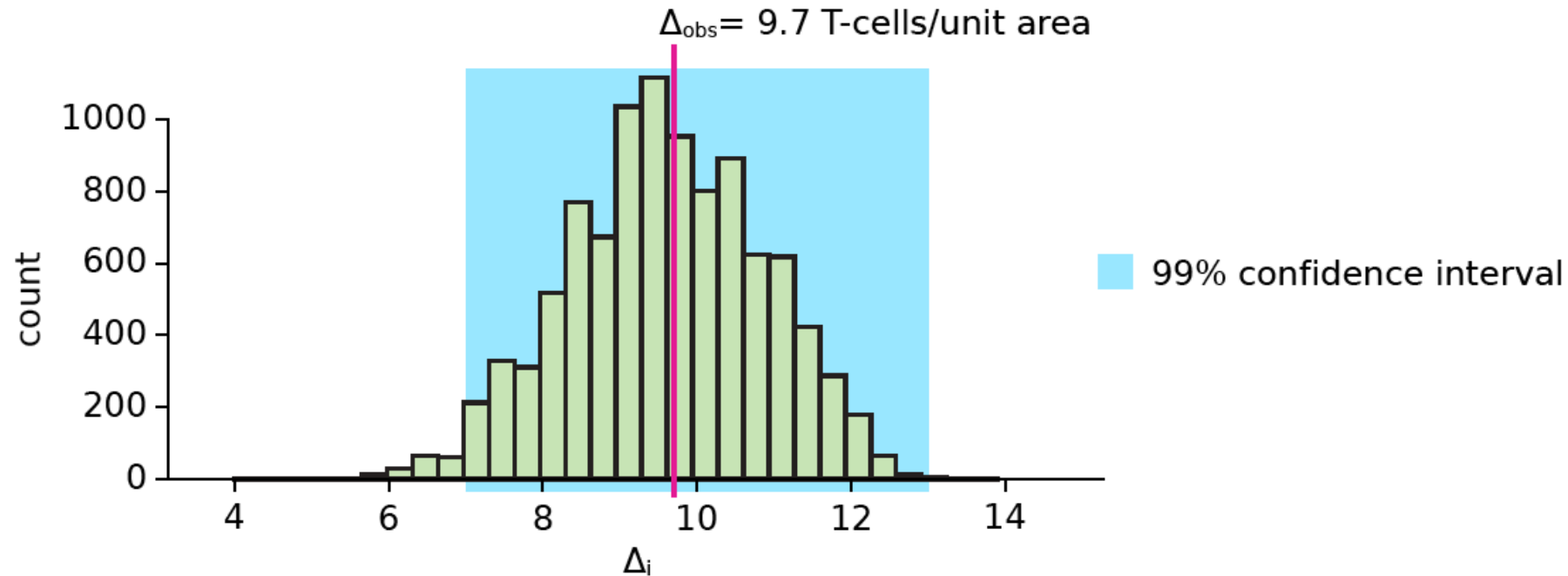
99% of differences from
resamples

Δ_H

top
0.5%

How to say it in a paper

“The Treatment group had higher median T-cell counts/unit area compared to Control; the difference in median counts was 9.7 (99% confidence interval (7, 13)) ($p < 0.001$)”



There is a formula for the Confidence Interval for the mean of a Normally-distributed (“bell-shaped”) distribution

We want

- “freedom of measure” (medians, etc.)
- arbitrary distributions

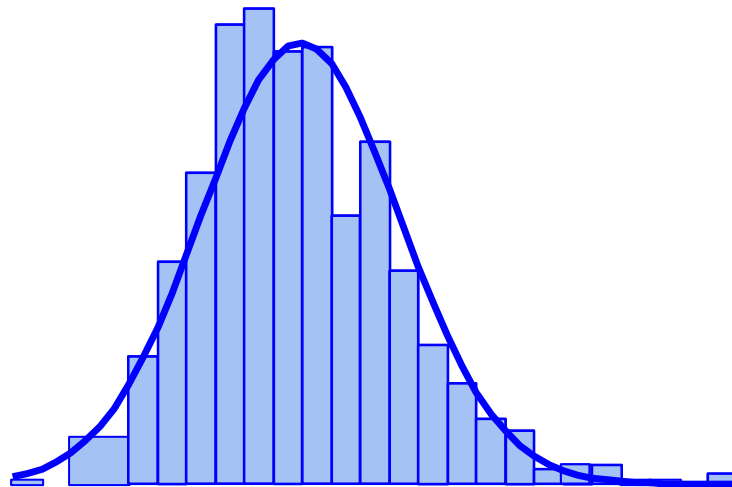
Finding a confidence interval
around any measure M ,
with no assumptions about
distributions



Visualizing Data

Why and How

as opposed to “Means and
Standard Deviations”



#1: Always make a picture of the data!

- **As the primary appreciation of the data set**
- **in order to see what types of numerical analyses are appropriate**

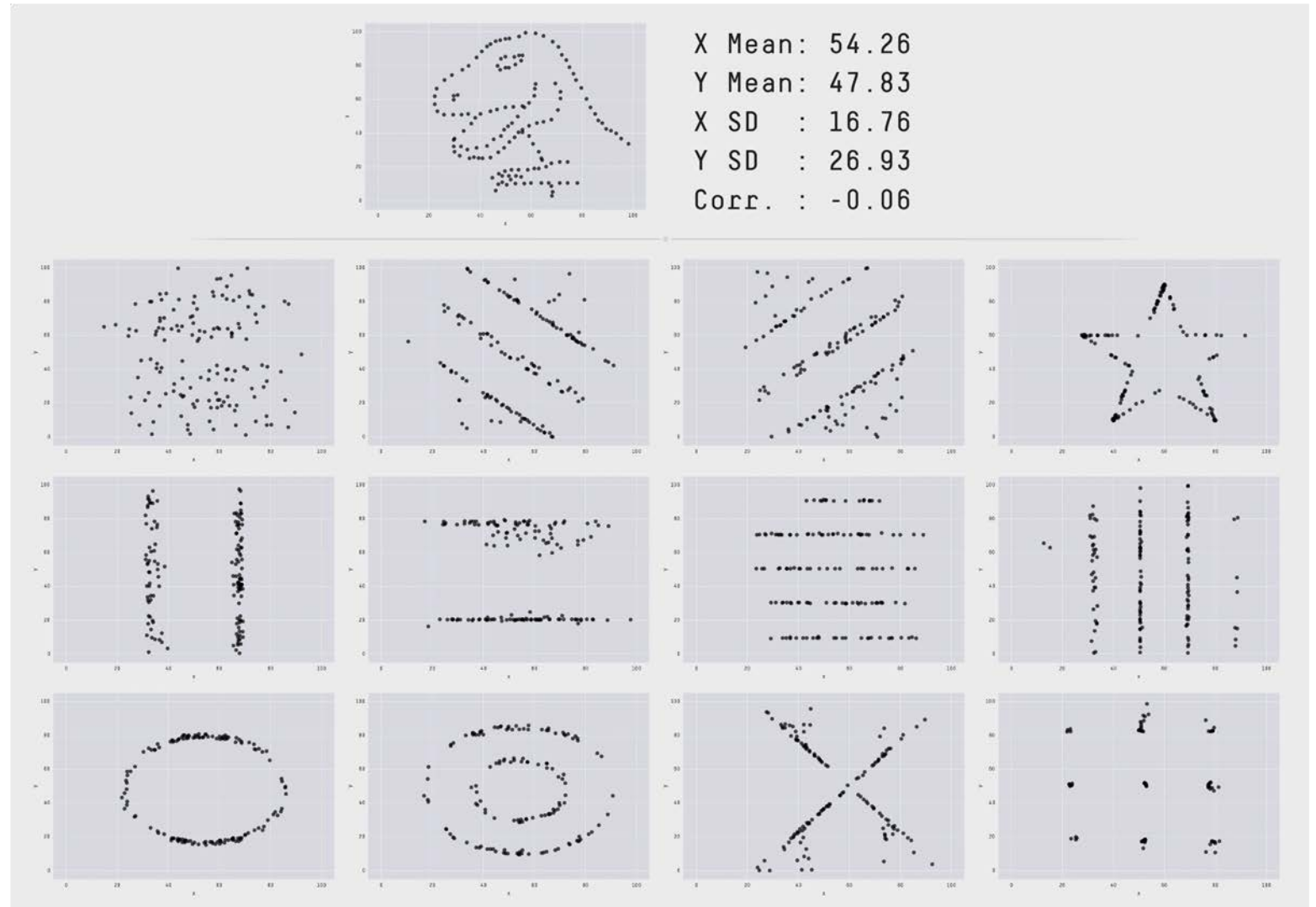
Always Visualize Your Data Before Analyzing!

Datasaurus

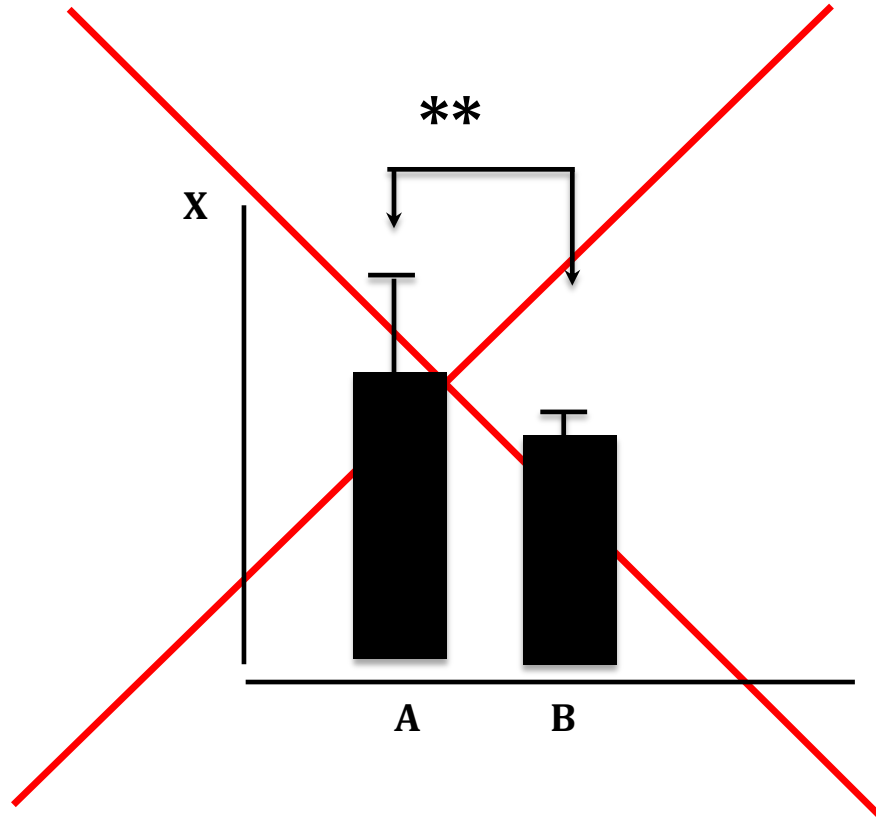
Dozen

By [Alberto Cairo](#)

(U Miami)



How Not to Present Data!!



- Hides the 'n'
- Hides the measure
- Hides the meaning of the 'error bar'
- Hides the statistical test

Show the data, don't conceal them

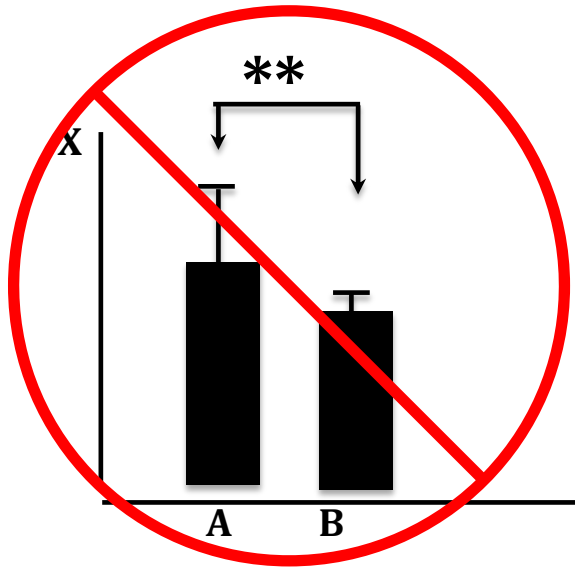
G. B. Drummond and S. L. Vowler

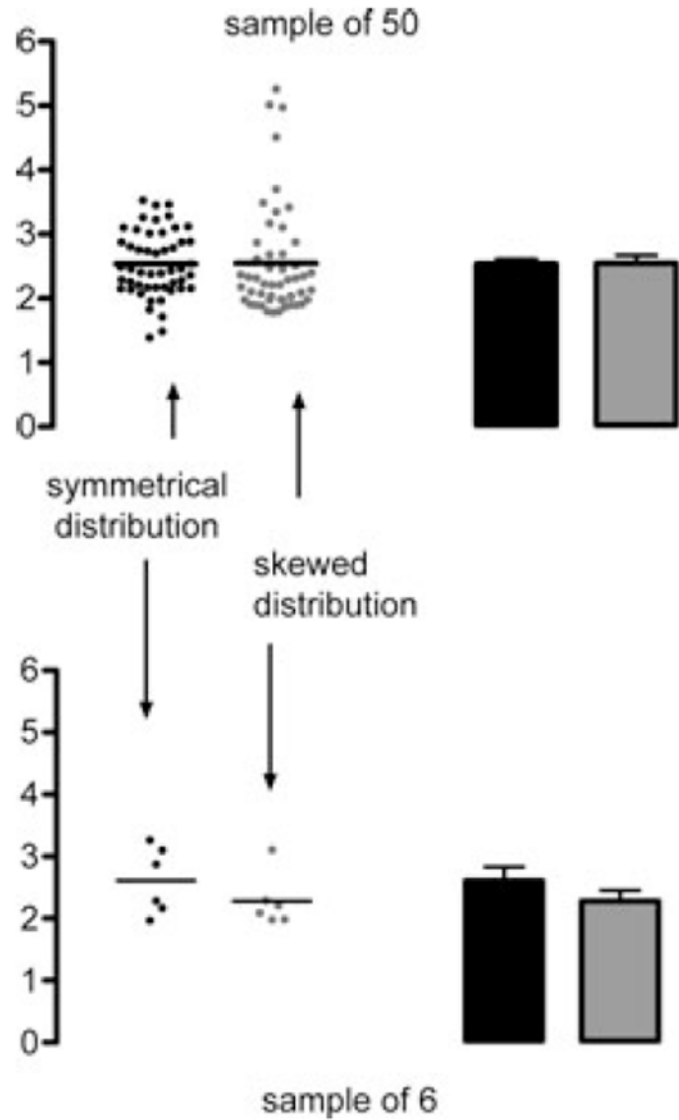
The Journal of
Physiology

Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*. Sarah Vowler is a medical statistician with Cancer Research UK.

“Dynamite plunger plots”

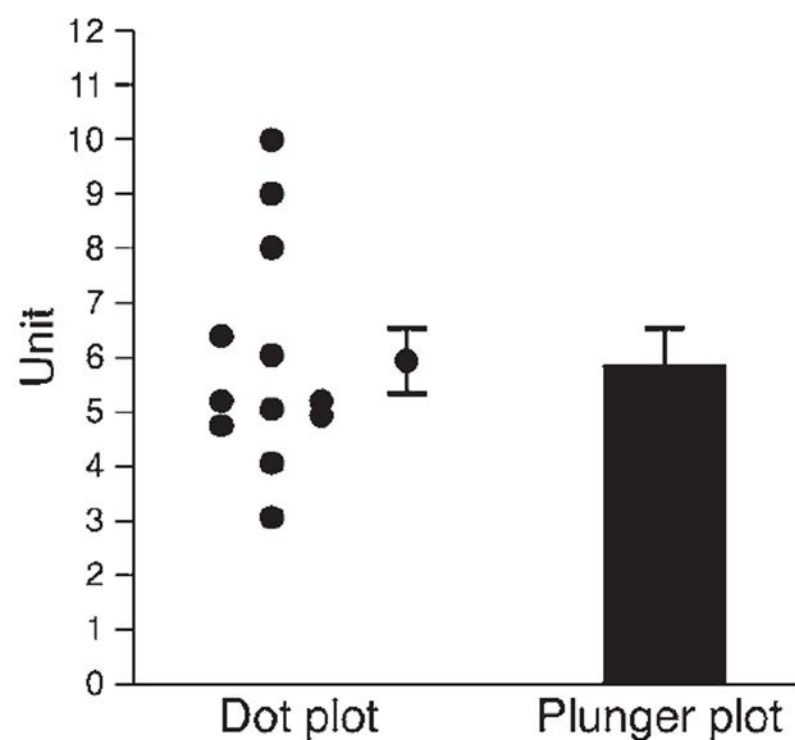
(Drummond and Vowler, *J Physiol* 2011)





“Dynamite plunger plots are never an appropriate way to plot the data.”

Drummond and Vowler
J Physiol 2011

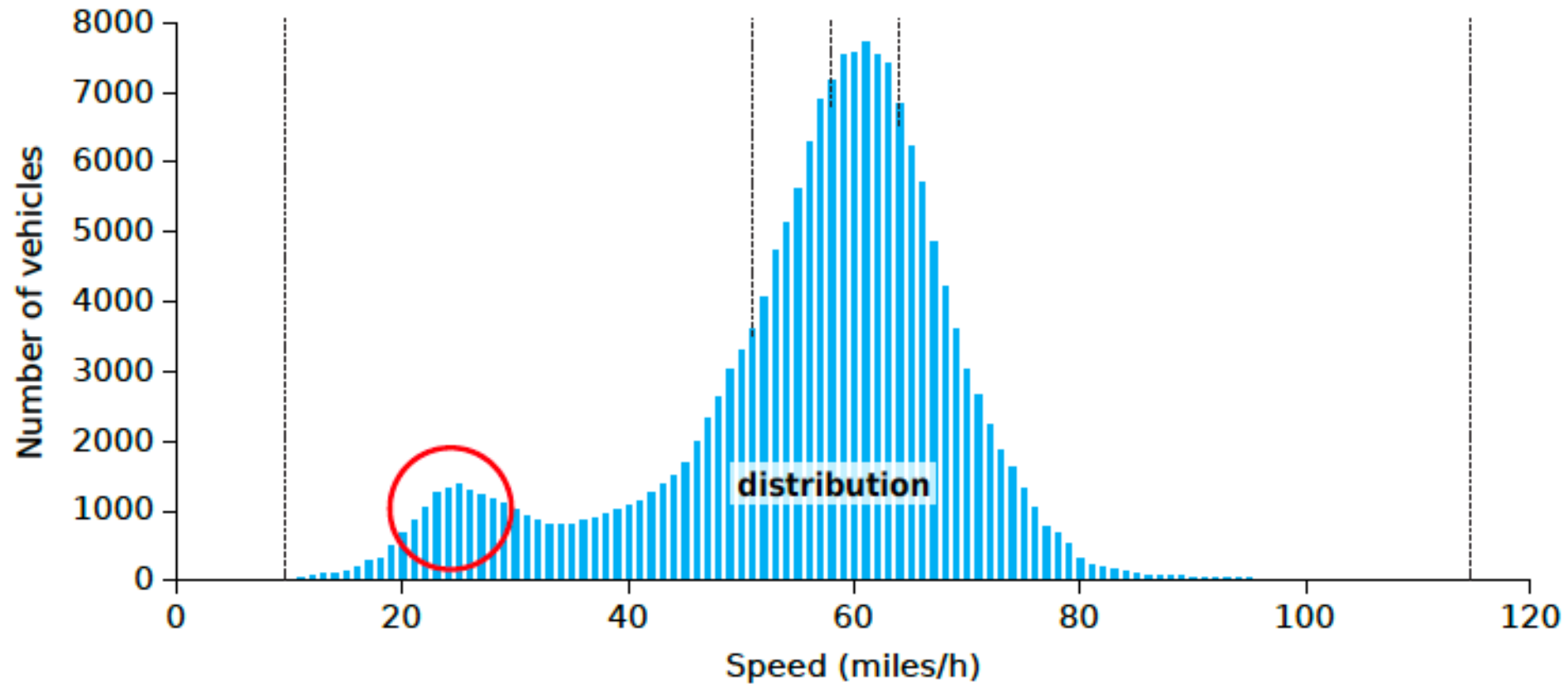


To maintain the highest level of trustworthiness of data, we are encouraging authors to display data in their raw form and not in a fashion that conceals their variance. Presenting data as columns with error bars (dynamite plunger plots) conceals data

Making a picture of the data distribution

- **to gain insight into the data**
- **To decide what statistics to use**

Vehicle speeds on a highway



AUGUST 1, 2022 | 4 MIN READ

How Florence Nightingale Changed Data Visualization Forever

The celebrated nurse improved public health through her groundbreaking use of graphic storytelling

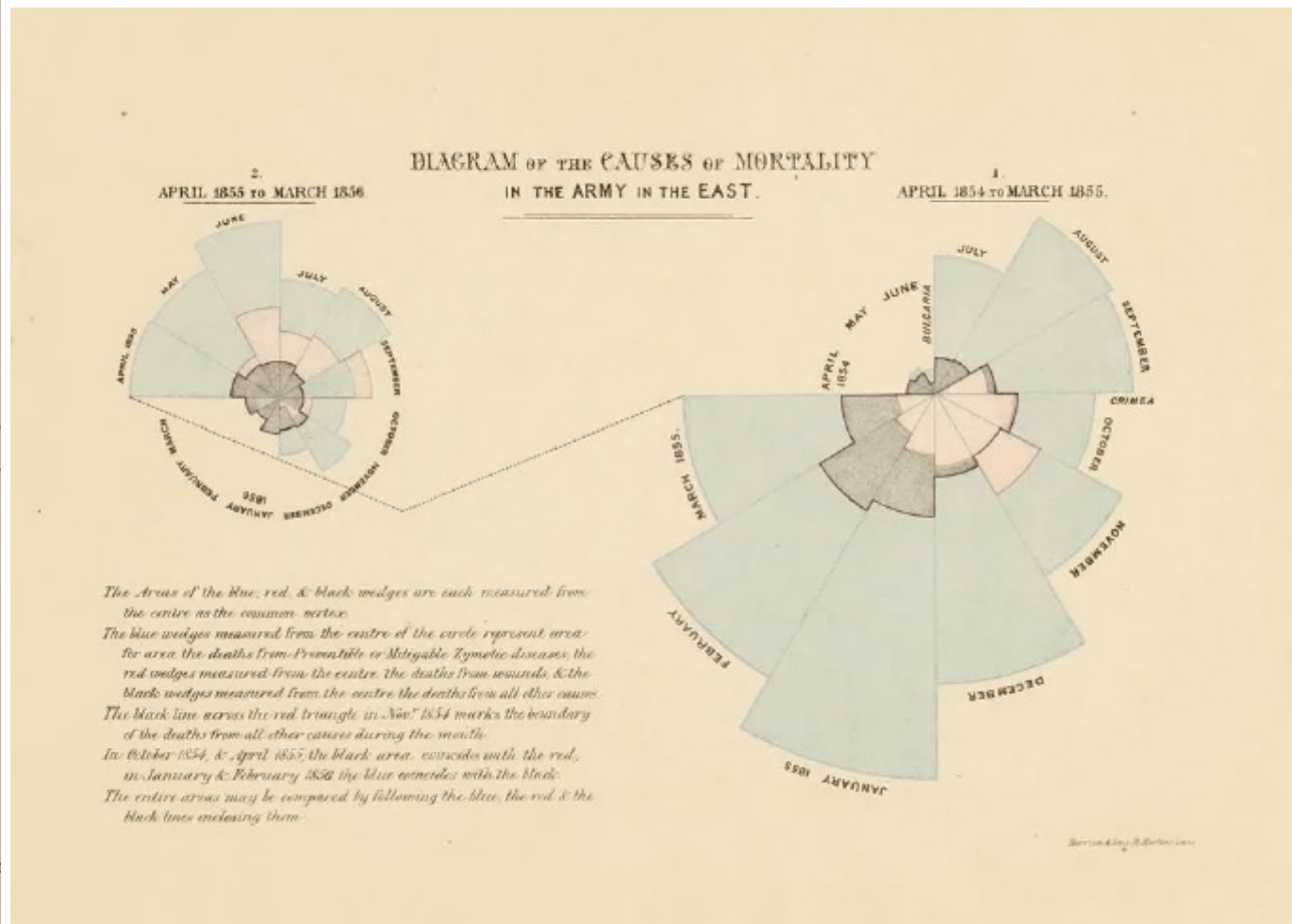
BY [RJ ANDREWS](#)



Killed and wounded, as returned on the 20 Sept. 1854.

Officers — Light Division										Men									
Killed					Wounded					Killed					Wounded				
Regt					Surg					Surg					Surg				
7th Fus					Surg					Surg					Surg				
1	—	4	—	3	—	4	—	11		38	—	80	—	70	—	22	—	172	
19 Regt	1	—	1	—	4	—	6	—	11	38	—	23	—	112	—	73	—	208	
23 rd	8	—	1	—	3	—	1	—	5	45	—	30	—	70	—	53	—	153	
33 rd	1	—	2	—	3	—	1	—	6	52	—	57	—	73	—	89	—	219	
77 th	0	—	0	—	0	—	0	—	0	2	—	5	—	7	—	4	—	16	
88 th	0	—	0	—	0	—	0	—	0	4	—	2	—	4	—	10	—	16	
2 Rifles	0	—	0	—	1	—	0	—	1	9	—	14	—	14	—	12	—	40	
3 Bkny	1	—	0	—	0	—	0	—	0	2	—	0	—	1	—	4	—	5	
6 Troop	0	—	0	—	0	—	0	—	0	1	—	0	—	0	—	0	—	0	
12					8					14					12				
20 th Dec					191					211					351				
Staff	0	—	0	—	1	—	4	—	5	81	—	89	—	178	—	149	—	416	
Regt	8	—	7	—	9	—	13	—	29	41	—	45	—	176	—	201	—	422	
Gen Staff	0	—	0	—	2	—	0	—	2	2	—	0	—	2	—	10	—	12	
1 st Bn	1	—	4	—	8	—	5	—	17	0	—	0	—	0	—	0	—	0	
3 rd Bn	0	—	0	—	0	—	0	—	0	1	—	0	—	1	—	0	—	1	
4 th Bn	0	—	0	—	1	—	0	—	1	0	—	0	—	0	—	0	—	0	
Canby	0	—	0	—	0	—	0	—	0	0	—	0	—	0	—	0	—	0	
1 st Regt	0	—	0	—	0	—	0	—	0	0	—	0	—	0	—	0	—	0	
Engineers	0	—	0	—	1	—	0	—	1	0	—	0	—	0	—	0	—	0	
21					19					36					34				
89					316					349					709				
1687																			

Note
 This is the original rough note made from the Regimental Returns that were sent in to me by the different Surgeons, but the corrected returns were subsequently found to be 25 Officers & 327 men killed, and 73 Officers & 1539 men wounded of which number 8 Surgeons accounts only for 1143, as you will see in Return No 2. In Return 3 he gives the result of the different operations up to the 27 Nov. 1854 and after that time I have no correct data in my possession that I can send you - but the number of wounded from that date up to March 55 was 5946 -





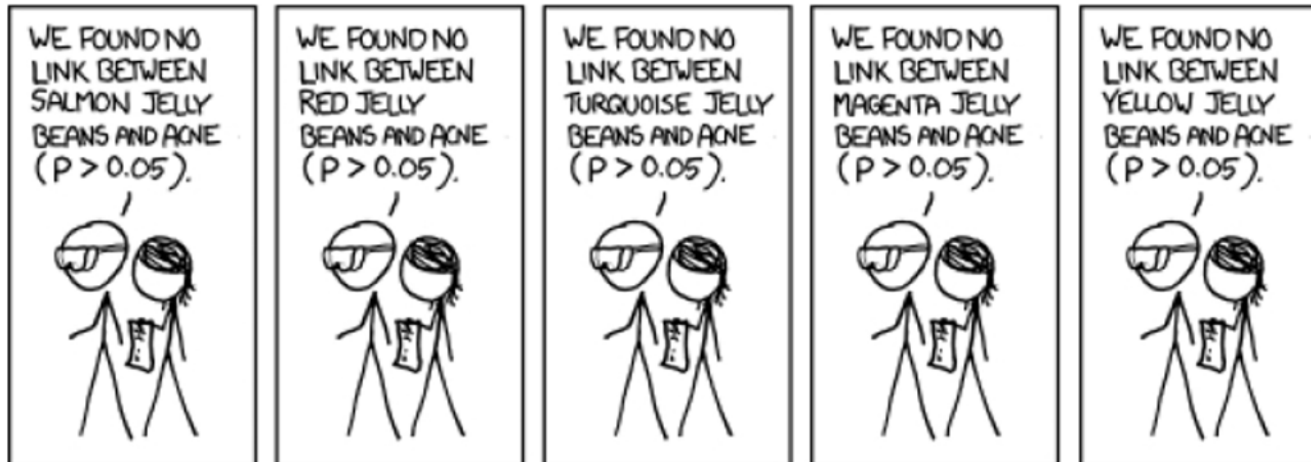
“P-hacking”

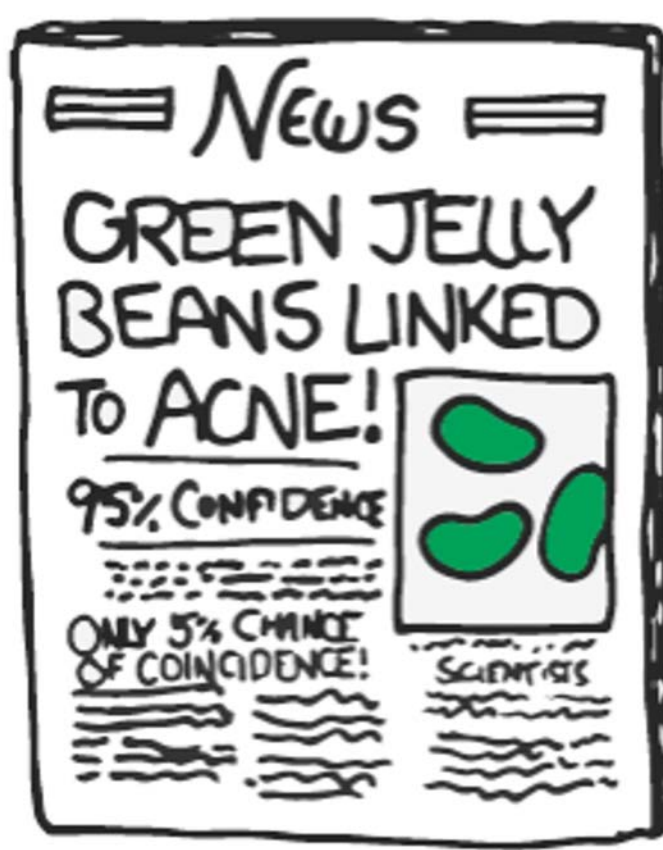
= Do lots and lots of tests and then publish the ones that turn out to be positive.

Green Jelly Beans

A scientist asks the lab to investigate whether there is any link between jelly beans and acne

boss: “OK, what about specific flavors of jelly beans?”





Q: What happens when we do 20 tests on random numbers with a “ $p < .05$ ” criterion for significance?



**A: on average,
there will be 1 False
Positive**

“p-hacking”

Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health

Peter C. Austin^{a,b,c,*}, Muhammad M. Mamdani^{a,d}, David N. Juurlink^{a,e,f}, Janet E. Hux^{a,c,e,f}

^aInstitute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5 Canada

^bDepartment of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada

^cDepartment of Health Policy, Management and Evaluation, University of Toronto, Canada

^dFaculty of Pharmacy, University of Toronto, Canada

^eClinical Epidemiology and Health Care Research Program (Sunnybrook & Women's College Site), Canada

^fDivision of General Internal Medicine, Sunnybrook & Women's College Health Sciences Centre and the University of Toronto, Canada

Accepted 19 January 2006

**Journal of
Clinical
Epidemiology**

Objectives: To illustrate how multiple hypotheses testing can produce associations with no clinical plausibility.

Study Design and Setting: We conducted a study of all 10,674,945 residents of Ontario aged between 18 and 100 years in 2000. Residents were randomly assigned to equally sized derivation and validation cohorts and classified according to their astrological sign. Using the derivation cohort, we searched through 223 of the most common diagnoses for hospitalization until we identified two for which subjects born under one astrological sign had a significantly higher probability of hospitalization compared to subjects born under the remaining signs combined ($P < 0.05$).

Results: We tested these 24 associations in the independent validation cohort. Residents born under Leo had a higher probability of gastrointestinal hemorrhage ($P = 0.0447$), while Sagittarians had a higher probability of humerus fracture ($P = 0.0123$) compared to all other signs combined. After adjusting the significance level to account for multiple comparisons, none of the identified associations remained significant in either the derivation or validation cohort.

Conclusions: Our analyses illustrate how the testing of multiple, non-prespecified hypotheses increases the likelihood of detecting implausible associations. Our findings have important implications for the analysis and interpretation of clinical studies. © 2006 Elsevier Inc. All rights reserved.

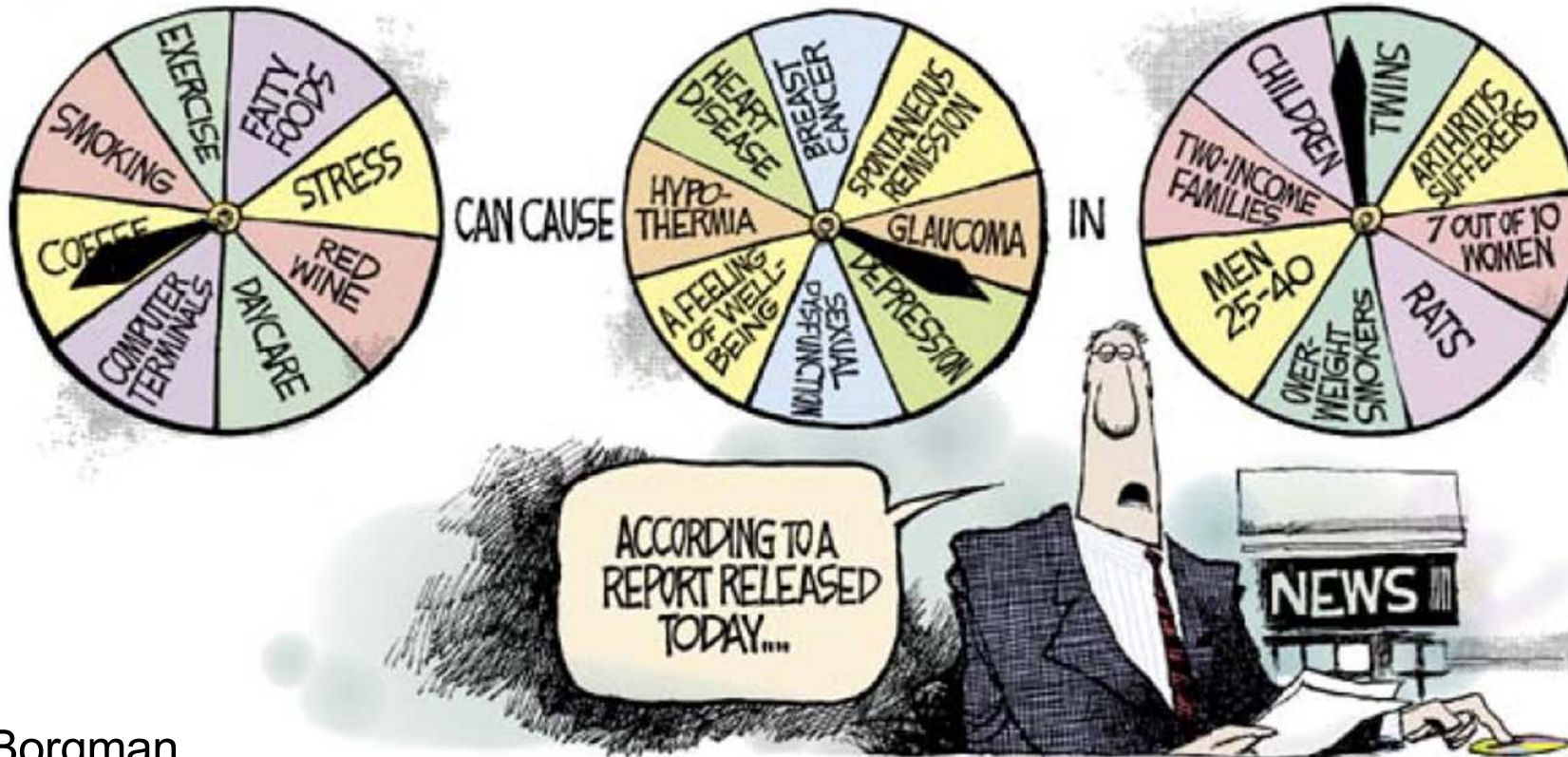


Compared 12 astrological signs X 223 diagnoses!

Today's Random Medical News

from the New England
Journal of
Panic-Inducing
Gobbledygook

JIM BORGMAN



Jim Borgman

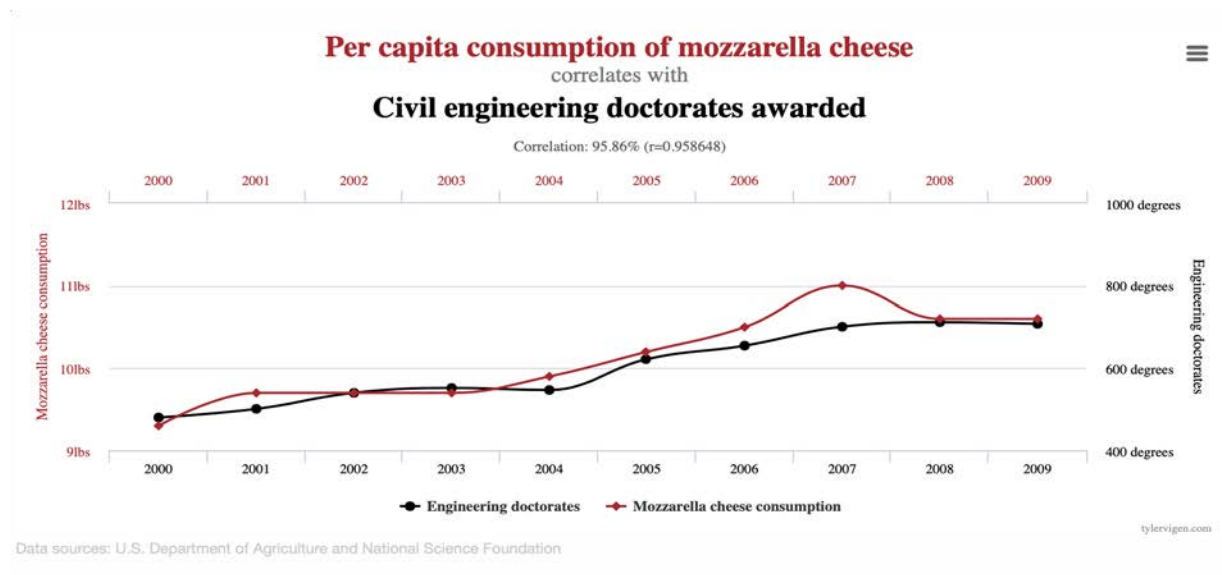
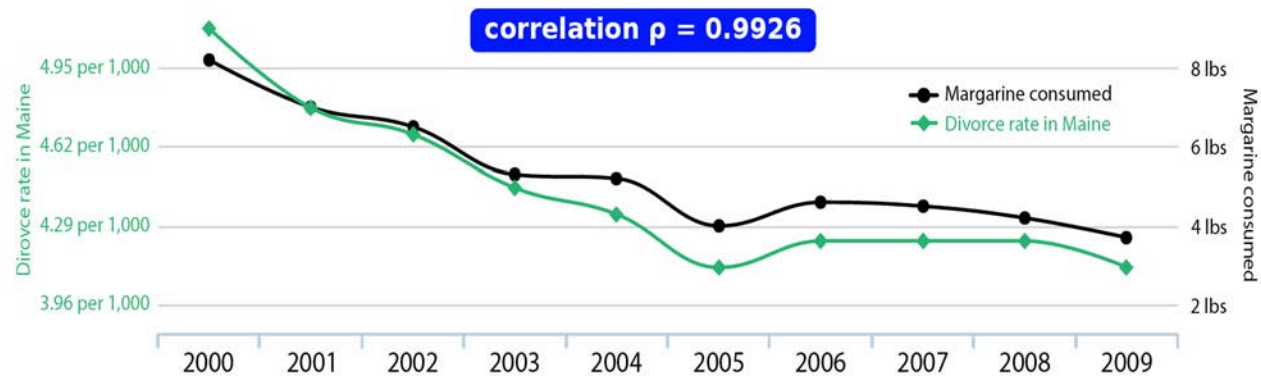
$8 \times 8 \times 8 = 512$ studies

5% of 512 $\cong$ 26

Therefore ~26 of these studies will produce False Positives at $p < .05$!

“data dredging” = p-hacking

Divorce rate in Maine correlates with Per capita consumption of margarine



great website!

tylervigen.com

Much Worse in Practice



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon:

An argument for multiple comparisons correction

Craig M. Bennett¹, Abigail A. Baird², Michael B. Miller¹, and George L. Wolford³

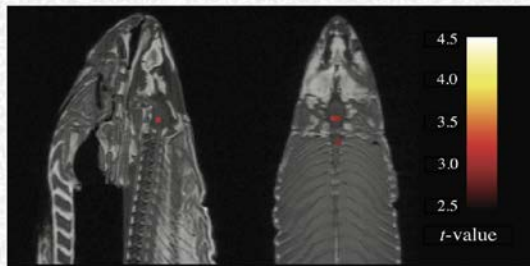
¹ Psychology Department, University of California Santa Barbara, Santa Barbara, CA;

² Department of Psychology, Vassar College, Poughkeepsie, NY;

³ Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

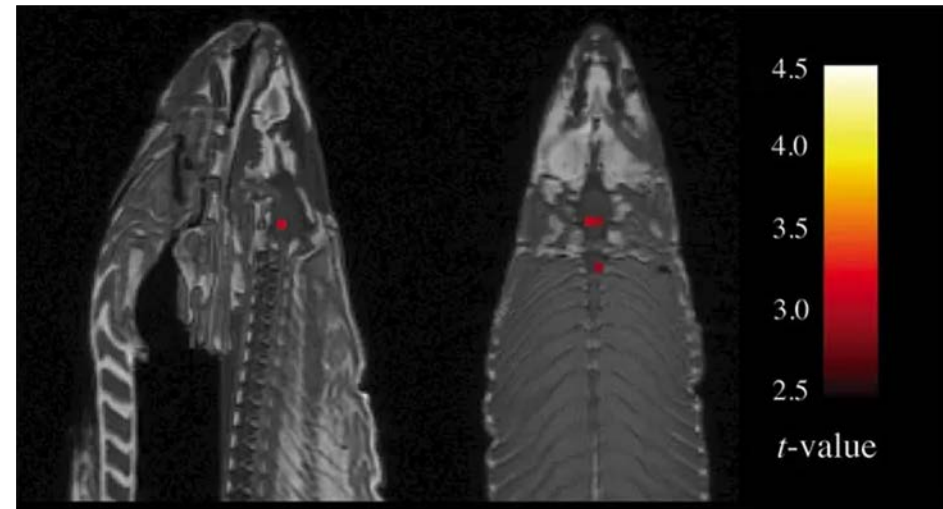
With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.



A t -contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon brain cavity (see above).

‘Higher’ standards, like $p < 0.001$, don’t help when we are doing 130,000 tests! -> need to correct for multiple tests!



Correction for Multiple Testing

We need to correct for multiple testing by reducing our false positive rate (α) for each pairwise test somehow.

We'll discuss two main methods:

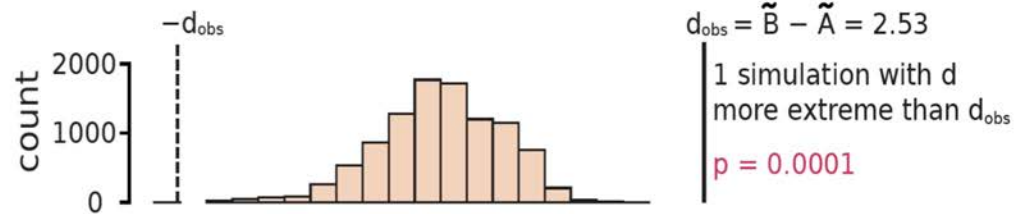
- Bonferroni
- Benjamini-Hochberg

⇒ Reduce statistical
significance & number
of false positives

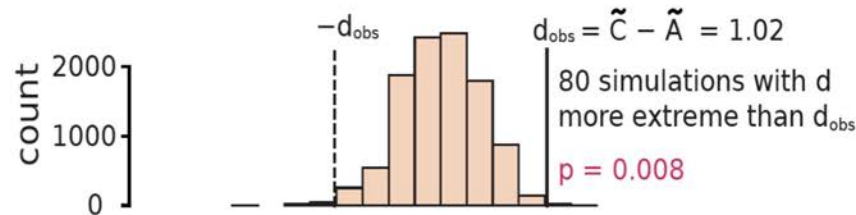
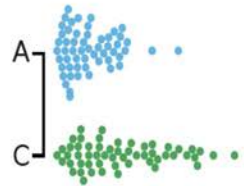
We call our desired α (uncorrected) the False Discovery Rate (FDR)

Benjamini-Hochberg for the T-cell data

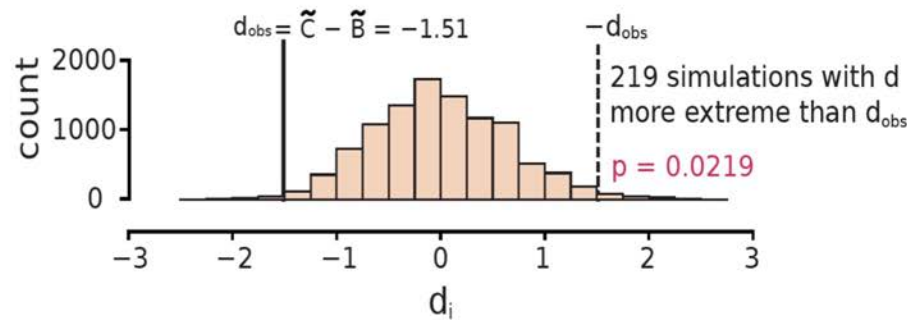
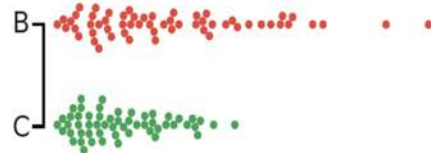
$m = 3$



$p_1 = 0.0001$

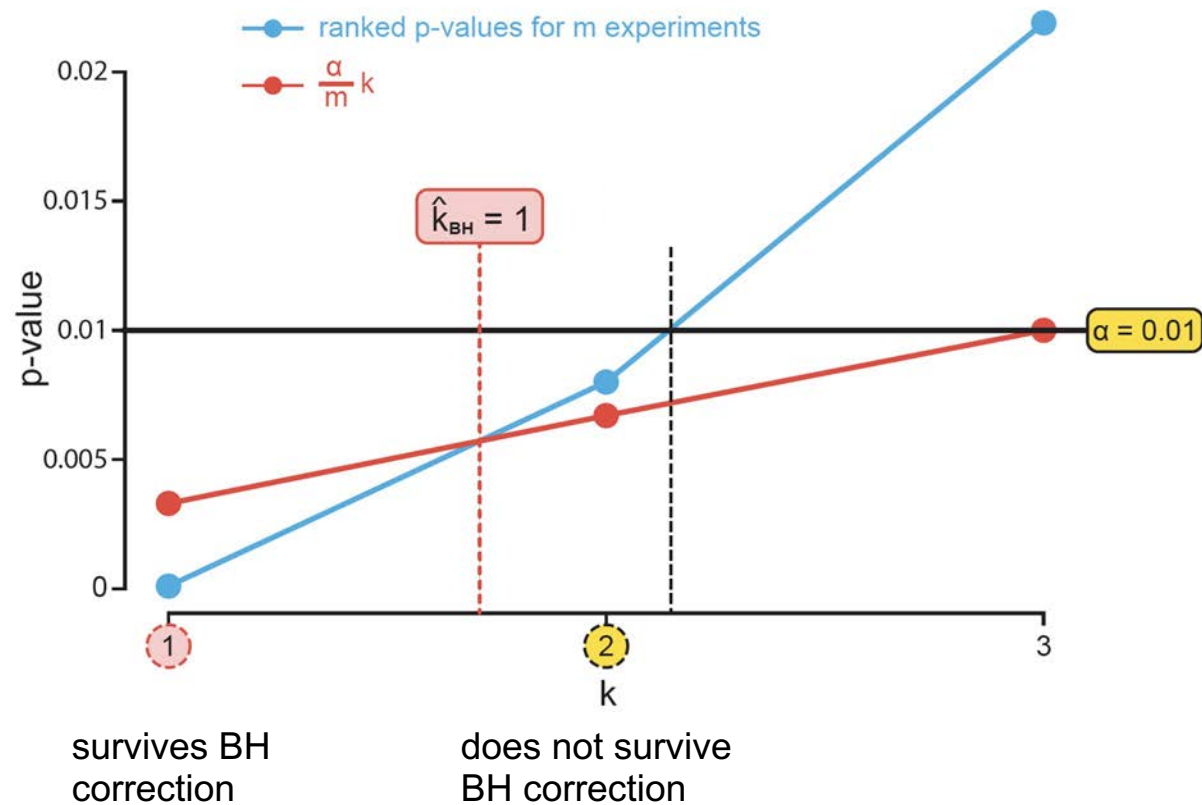


$p_2 = 0.008$



$p_3 = 0.0219$

①	②	③	④	⑤
k	p_k	$\alpha \cdot \frac{k}{m}$	is $p_k < \alpha$?	is $p_k < \alpha \cdot \frac{k}{m}$?
1	0.0001	0.0033	✓	✓
2	0.0080	0.0067	✓	✗ $\widehat{k}_{BH} = 1$
3	0.0219	0.0100	✗	✗





Is factor A **independent** of factor B?

Was survival on the Titanic **independent** of ticket class?

	First Class	Steerage
died	122	1368
survived	203	508

Was hospitalization/death **independent** of whether the subject took the drug?

	Hospitalized or died?	
	Yes	No
placebo	53	324
drug	28	357

Is the likelihood of a heart attack **independent** of diabetes status?

2 X 2 Contingency table

		heart attack?	
		yes	no
diabetes?	yes	n_1	n_2
	no	n_3	n_4

“affected by” or “independent of” are really questions about Comparing Proportions

Titanic: Is the **proportion** of survivors in First Class different from the proportion of survivors in steerage?

	First Class	Steerage
died	122	1368
survived	203	508

drug trial: was the **proportion** of hospitalized/died people in the placebo class the same as in First Class?

	Hospitalized or died?	
	Yes	No
placebo	53	324
drug	28	357

diabetes/heart disease: Is the **proportion** of heart attacks the same in diabetics as in non-diabetics?

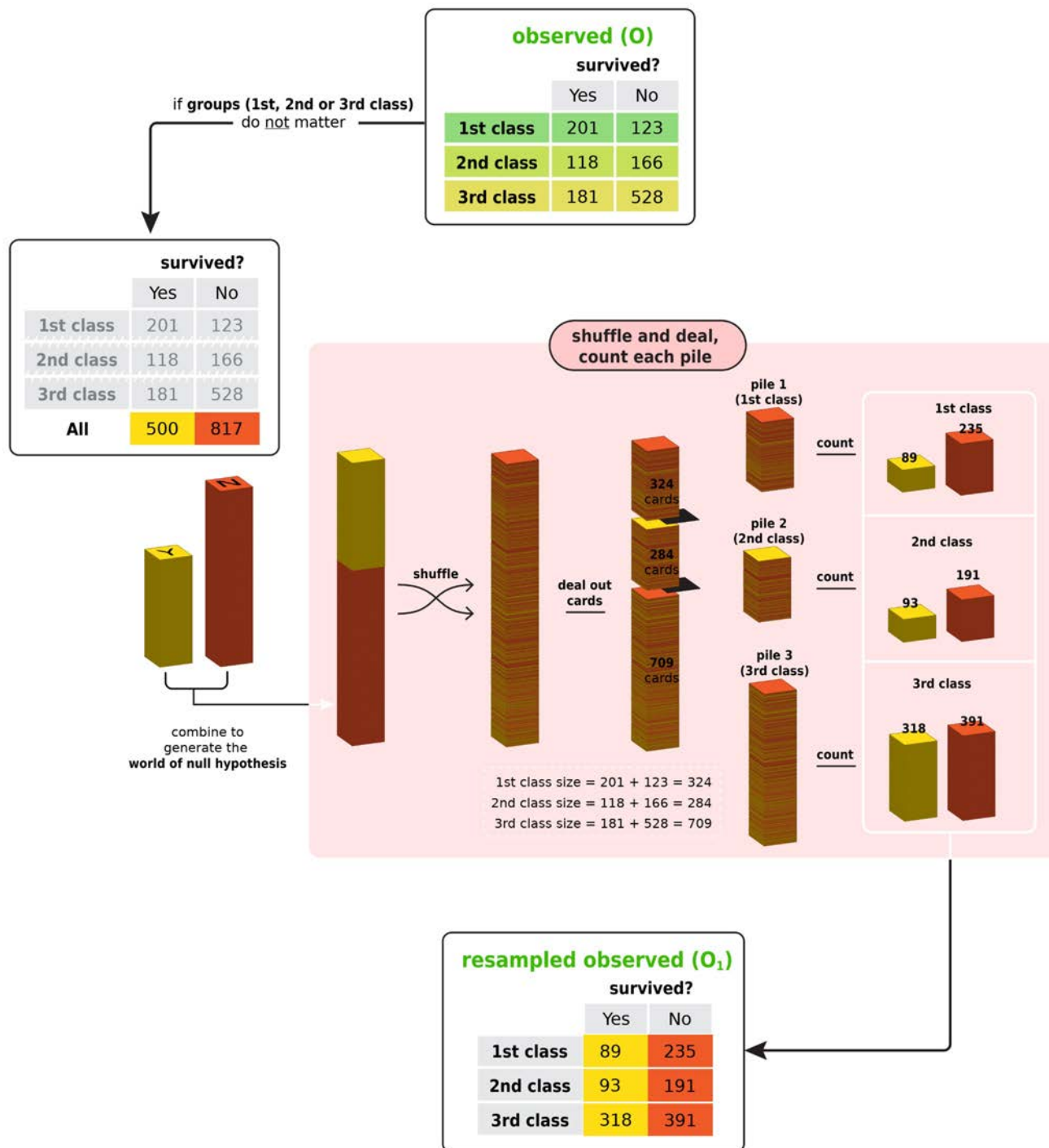
diabetes?

	heart attack?	
	yes	no
yes	n_1	n_2
no	n_3	n_4

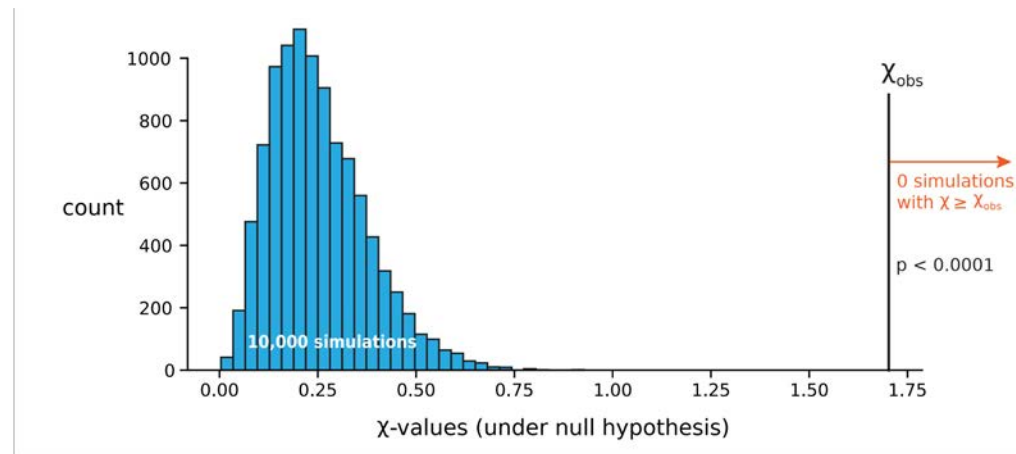
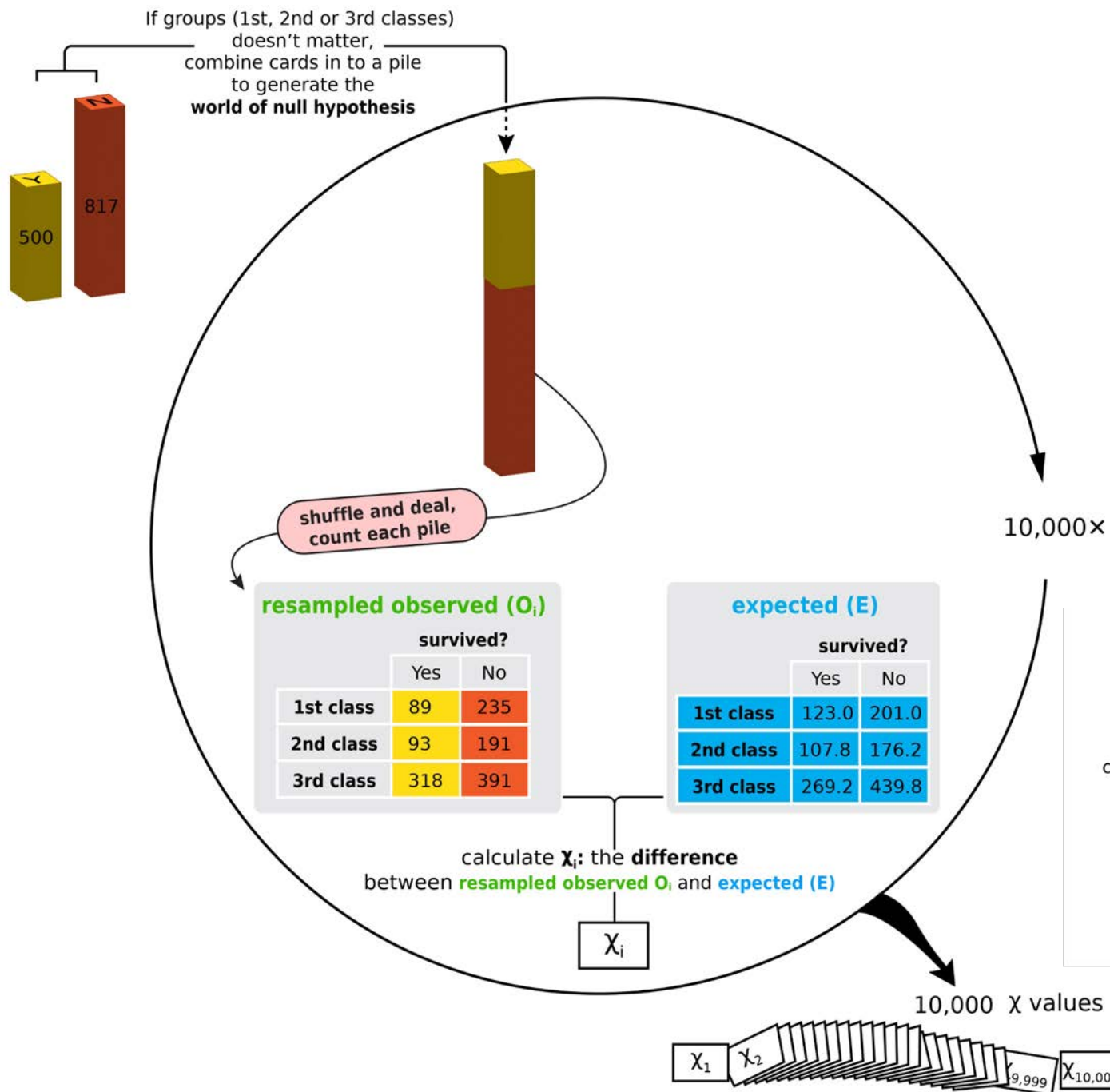
Survival on the Titanic

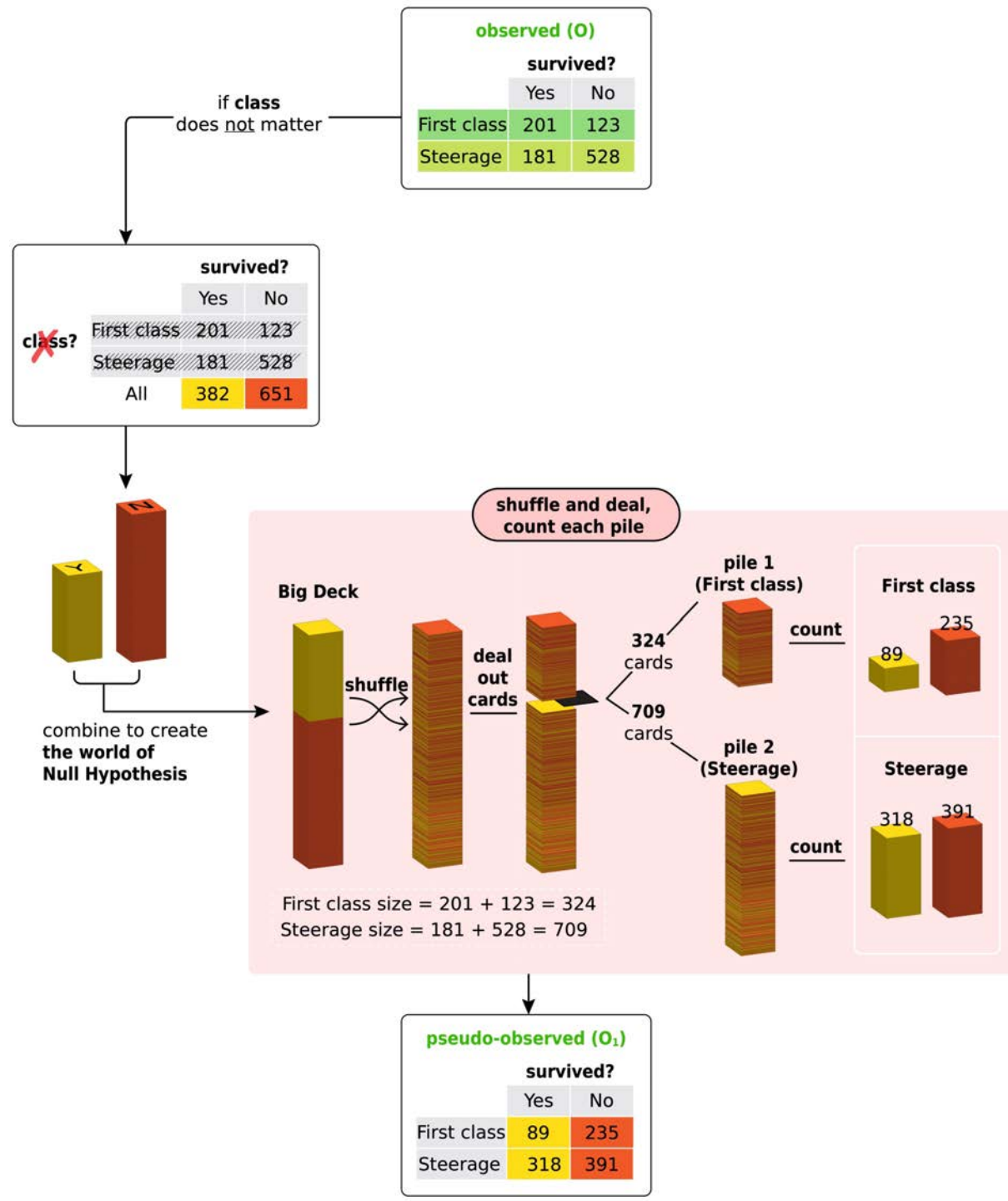


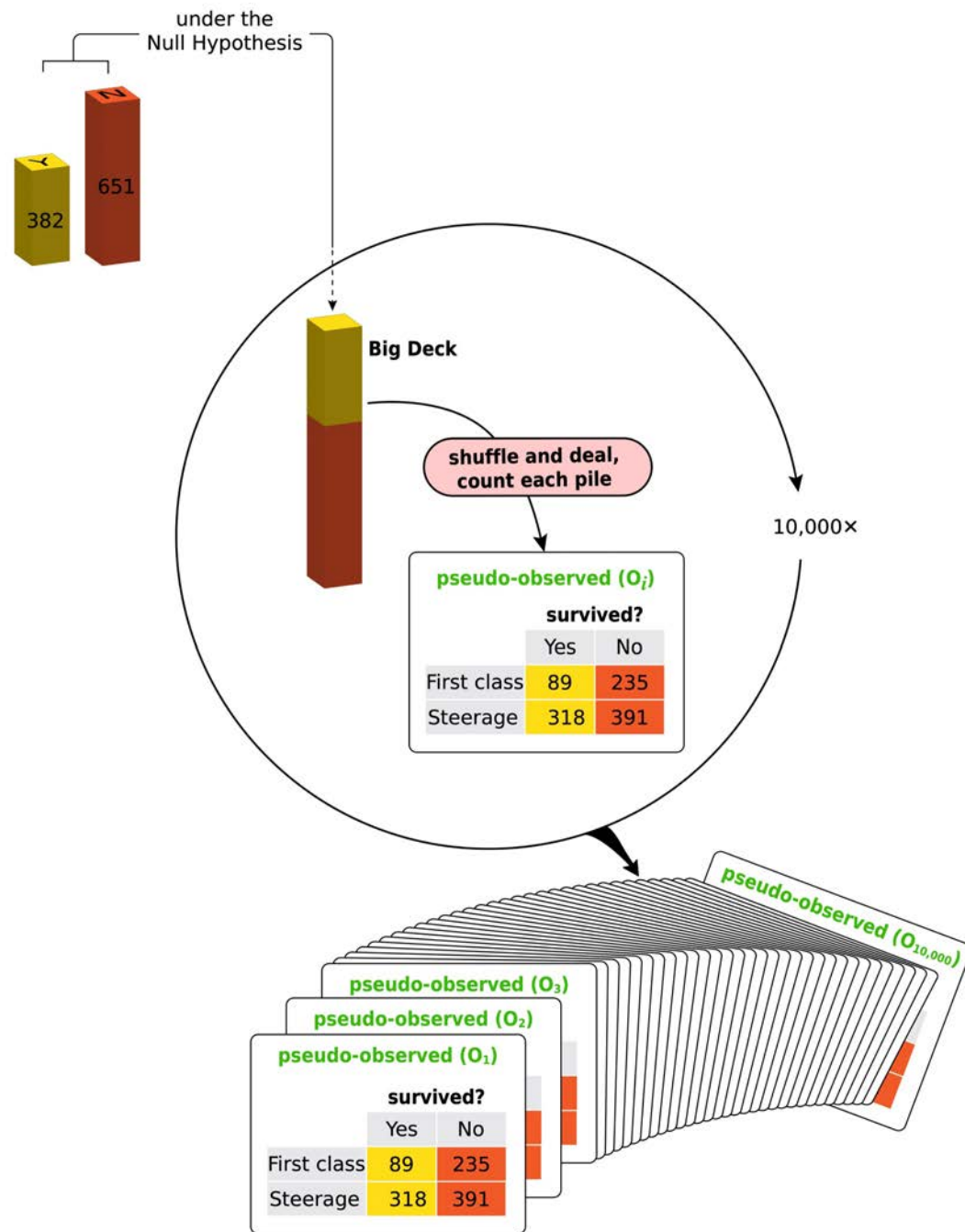
	Class (predictor)		Total
Outcome	First class	Steerage	
Survived	203	508	711
Died	122	1368	1490
Total	325	1876	2201



- You have Y number of 'Life' cards, and N number of 'Death' cards
- Shuffle the deck and deal them out into 3 piles of size (1st class, 2nd class and 3rd class)
- how often did you see a $|x|$ this large?







Relative Risk

Proportion of died in drug-yes group = $c/(a+c)$

Proportion of died in drug-no group = $d/(b+d)$

Contingency table:

		took drug?	
		yes	no
survived?	yes	a	b
	no	c	d

$$\text{Relative risk of dying due to taking drug} = \frac{\text{Proportion of died in drug-yes group}}{\text{Proportion of died in drug-no group}}$$

$RR = 0.7 \Rightarrow$ People who took drug had 70% of the risk of dying that people who didn't take the drug had.

Relative Risk

$$\text{Relative risk of outcome due to treatment} = \frac{\text{Proportion (Probability) of outcome in group with treatment}}{\text{Proportion (Probability) of outcome in group without treatment}}$$

- If treatment reduces probability of outcome, $RR < 1$
- If treatment doesn't affect probability of outcome, $RR = 1$
- If treatment increases probability of outcome, $RR > 1$

ORIGINAL ARTICLE

TRILOGY-ACS

Prasugrel versus Clopidogrel for Acute Coronary Syndromes without Revascularization

RESULTS

At a median follow-up of 17 months, the primary end point of death from cardiovascular causes, myocardial infarction, or stroke among patients under the age of 75 years occurred in 13.9% of the prasugrel group and 16.0% of the clopidogrel group (hazard ratio in the prasugrel group, 0.91; 95% confidence interval [CI], 0.79 to 1.05; P=0.21). Similar results were observed in the overall population. The pre-

“hazard ratio” = Relative Risk

Q: Calculating RR for Titanic

Calculate the **relative risk of dying** due to being in first-class

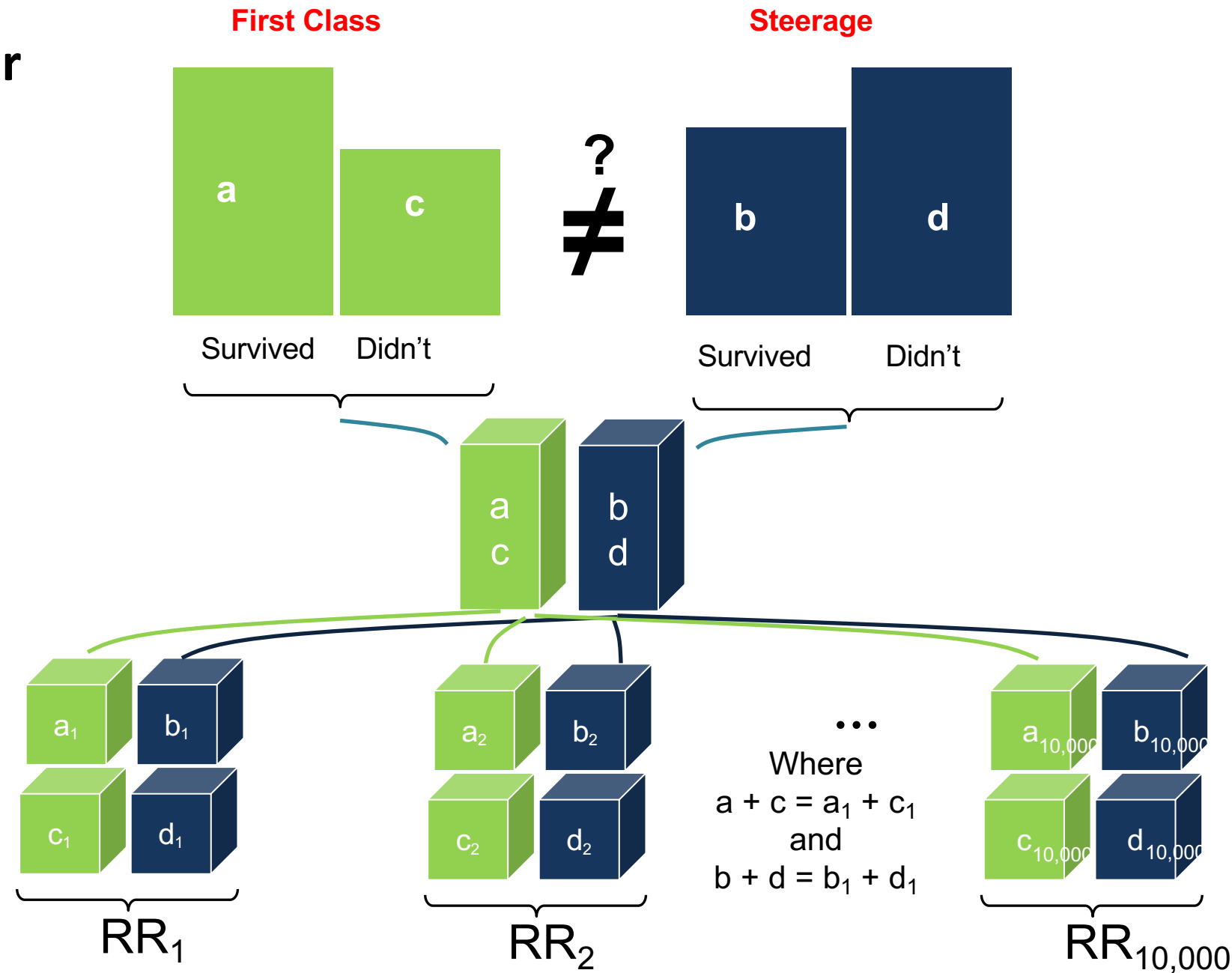
	Class (predictor)		
Outcome	First class	Steerage	Total
Survived	203	508	711
Died	122	1368	1490
Total	325	1876	2201

Relative Risk of death in 1st class = $\frac{\text{proportion with outcome (death) in 1st class}}{\text{proportion with outcome (death) in steerage}}$

$$\begin{aligned} RR_{obs} &= \frac{122/(122 + 203)}{1368/(1368 + 508)} \\ &= 0.51 \end{aligned}$$

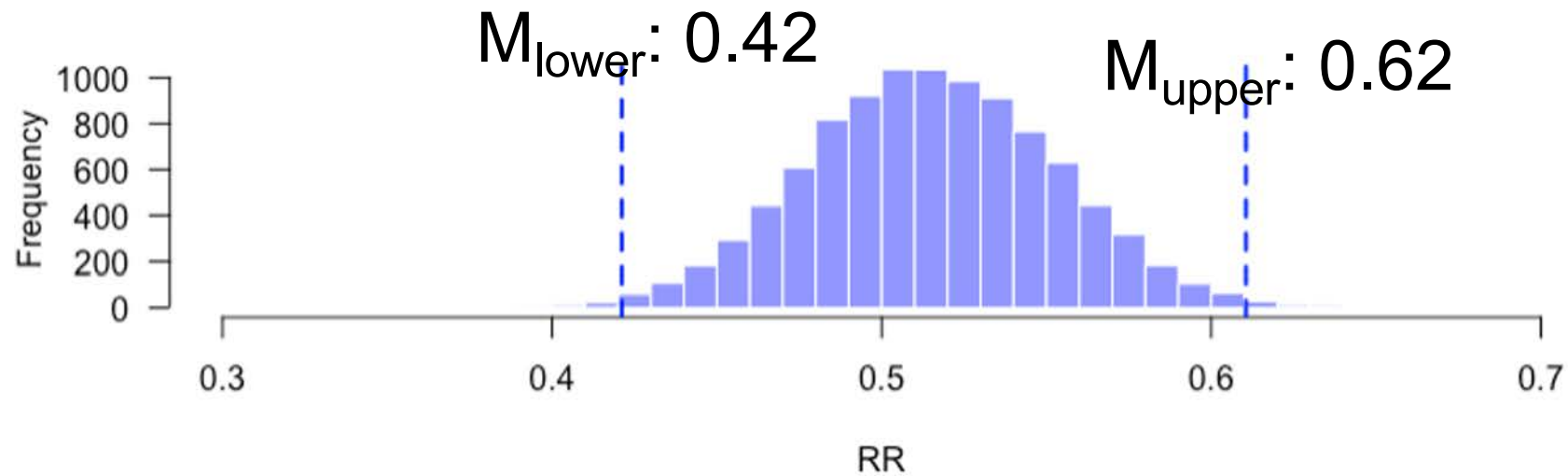


Confidence Interval for RR



Confidence Interval on RR for Titanic

The relative risk of dying in first class on the Titanic (compared to steerage) is 0.51.



Take home messages:

- visualize the data
- Choose scientifically and distributionally appropriate measures
- Never assume quantities are Normally Distributed
- Find p-values and Confidence intervals by resampling the data

Problem: The p-value is the answer to the wrong question.

p-values are an answer to the question:

how likely is this data (D) given the null hypothesis?

$$\textit{prob}(D | H_0)$$

But what we really want is:

$$\textit{prob}(H_0 | D)$$

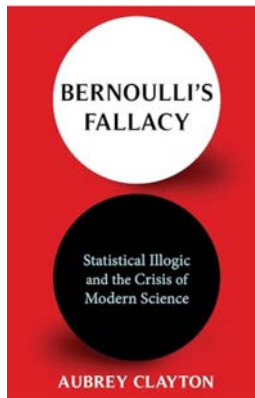
How likely is the null hypothesis given this data?

H_0 = Null Hypothesis D = your data

$$\textit{prob} (H_0 \mid D) \neq \textit{prob} (D \mid H_0)$$

what you really wanted

p-value



What does “ $p < .01$ ” really mean?

“Results this extreme happen less than 1% of the time (under the null hypothesis)”

“If you did this experiment many times and everything was random, you would get a result this extreme less than 1% of the time”

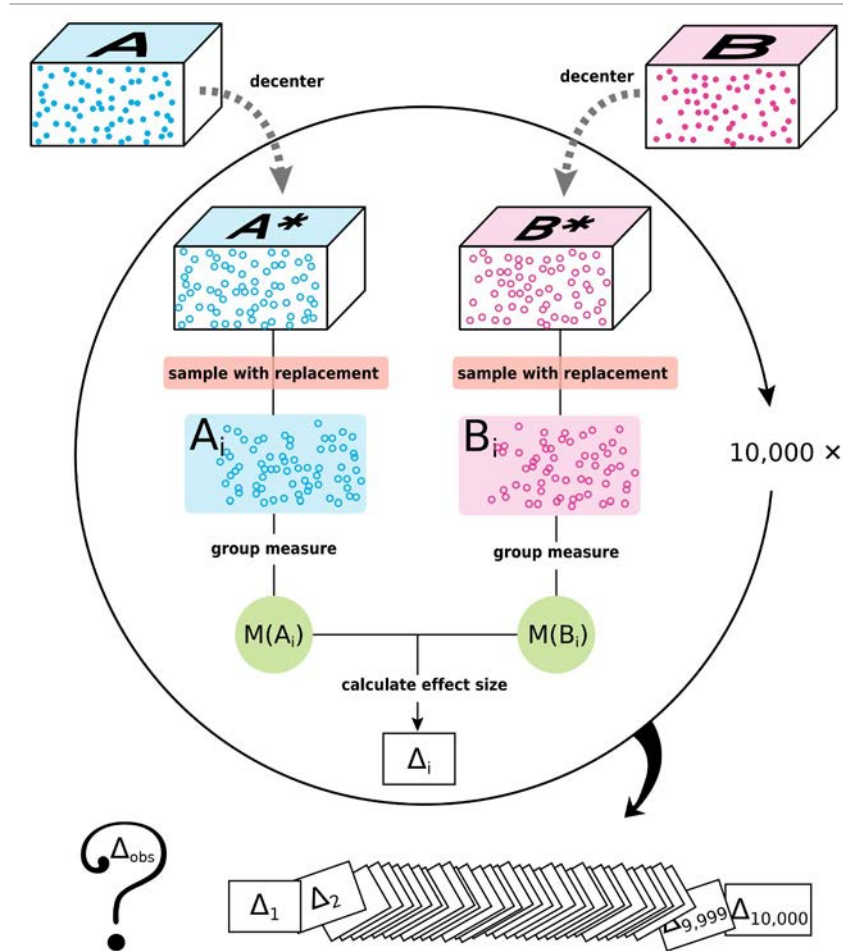
It does NOT mean “the probability of the null hypothesis being false is 1%”!!!

$$p(H_0|\text{data}) \neq p(\text{data}|H_0)$$



Two Box Re-centered

Null Hypothesis: A and B come from *different* underlying distributions (different variance, MADAM, skew etc.), *but with the same mean (or median) value*



“Welch’s t-test”

$$t = \left(\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \right)^{-1} \left(\frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{(n_A-1)\hat{\sigma}_A^2 + (n_B-1)\hat{\sigma}_B^2}{n_A+n_B-2}}} \right)$$

$$p(t \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu \pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu} \right)^{-\frac{\nu+1}{2}}$$

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

$$\nu = \frac{\left(\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B} \right)^2}{\left(\frac{\hat{\sigma}_A^4}{n_A^2 \nu_A} + \frac{\hat{\sigma}_B^4}{n_B^2 \nu_B} \right)^2}$$