R Markdown:

The Bigger Picture

April 4, 2019

Garrett Grolemund, RStudio

@StatGarrett

Amgen 2012

Amgen 2012: could reproduce only **6** of **53** landmark results

75%–90%

of preclinical results cannot
be reproduced (estimate)

Begley CG, Ioannidis JP (2015) Reproducibility in Science: improving the standard for basic and preclinical research.
Circ Res 116: 116–126. pmid:25552691

$28 billion

£ 22B € 24B

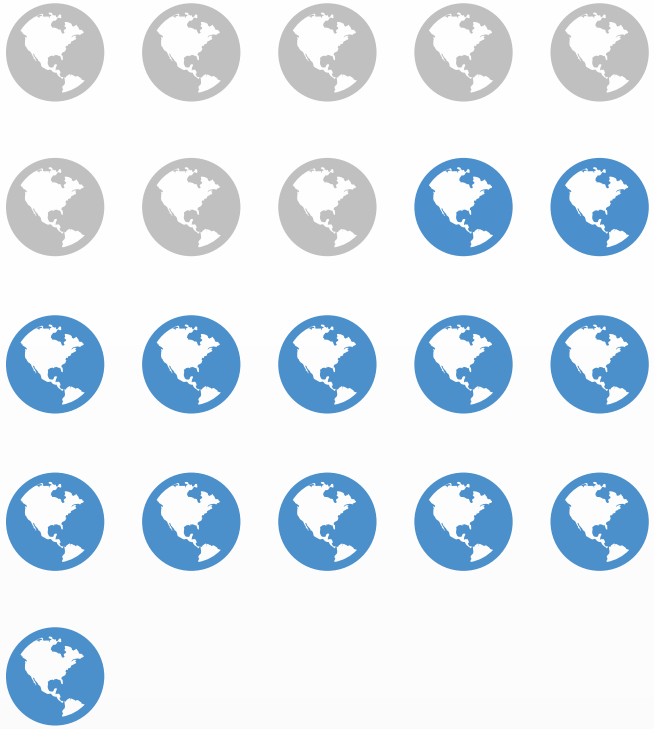
Costs to biomedical industry per year, US

Freedman LP, Cockburn IM, Simcoe TS (2015) The Economics of Reproducibility in Preclinical Research. PLoS Biol 13(6): e1002165. https://doi.org/10.1371/journal.pbio.1002165

Economics

6 of **18**

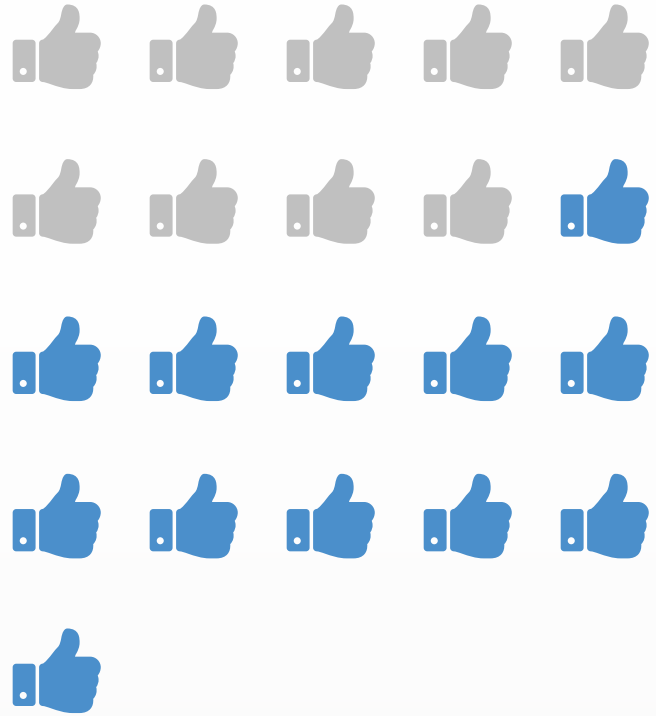
Nature & Science



13 of **21**

Coin Tosses

that were heads



12 of **21**

Why Do So Many Studies Fail to Replicate?

- The New York Times, May 2016

Scientists' Elusive Goal: Reproducing Study Results

- The Wall Street Journal, Dec 2011

Challenges in irreproducible research

- Nature, July 2018

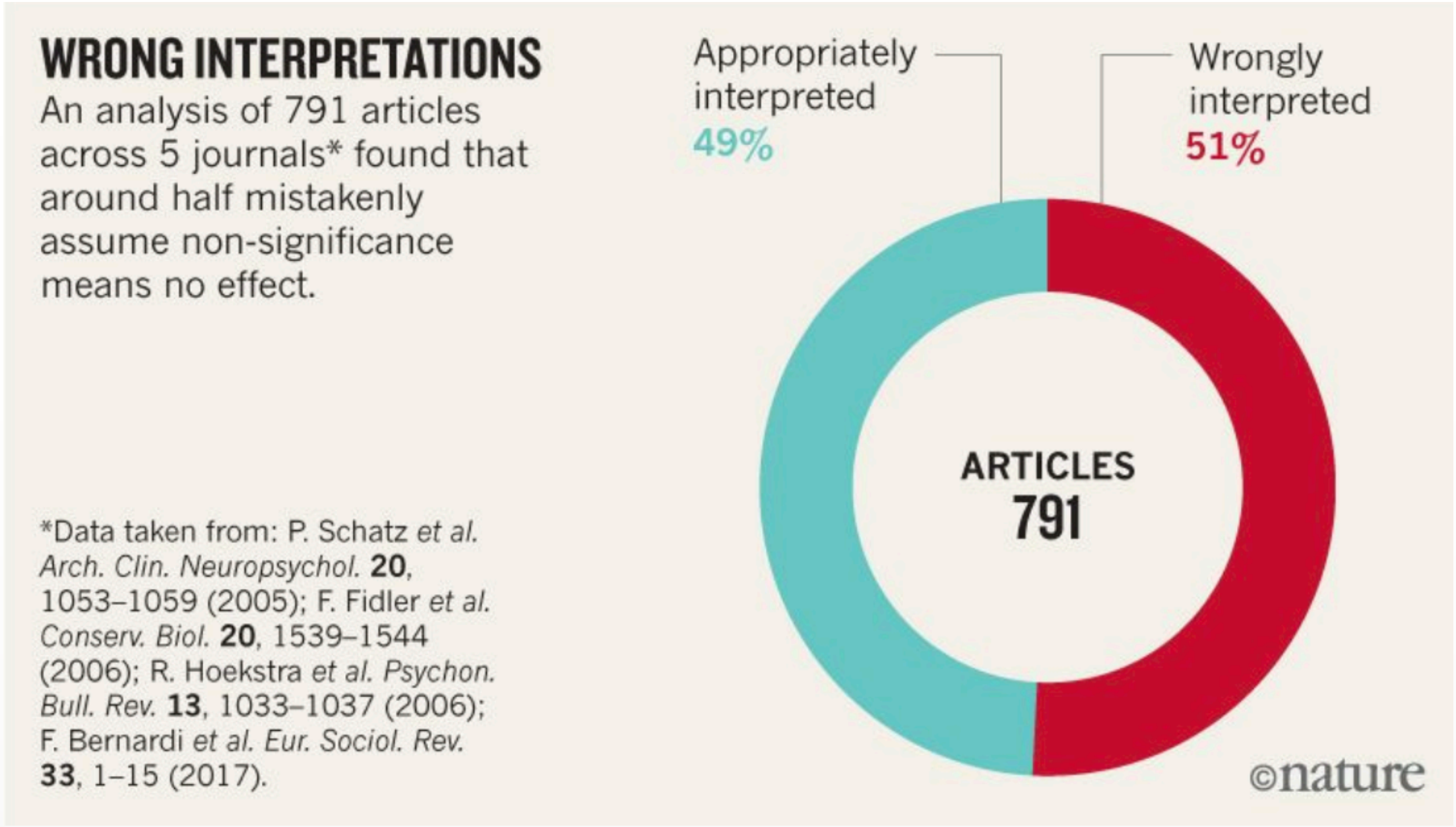
It is time to stop using the term "statistical significance" in its entirety.

Nor should variants such as "significantly different," "$p < 0.05$" and "nonsignificant" survive.

- The American Statistical Association , March 2019

THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION



. Scientists rise up against statistical significance. Nature. 2019.*www.nature.com/articles/d41586-019-00857-9*

In the case considered in Ex. 13·1, we have therefore

$$\frac{18!\,12!\,17!\,13!}{30!}\left\{\frac{1}{2!\,3!\,10!\,15!}+\frac{1}{2!\,11!\,16!}+\frac{1}{1!\,12!\,17!}\right\}$$

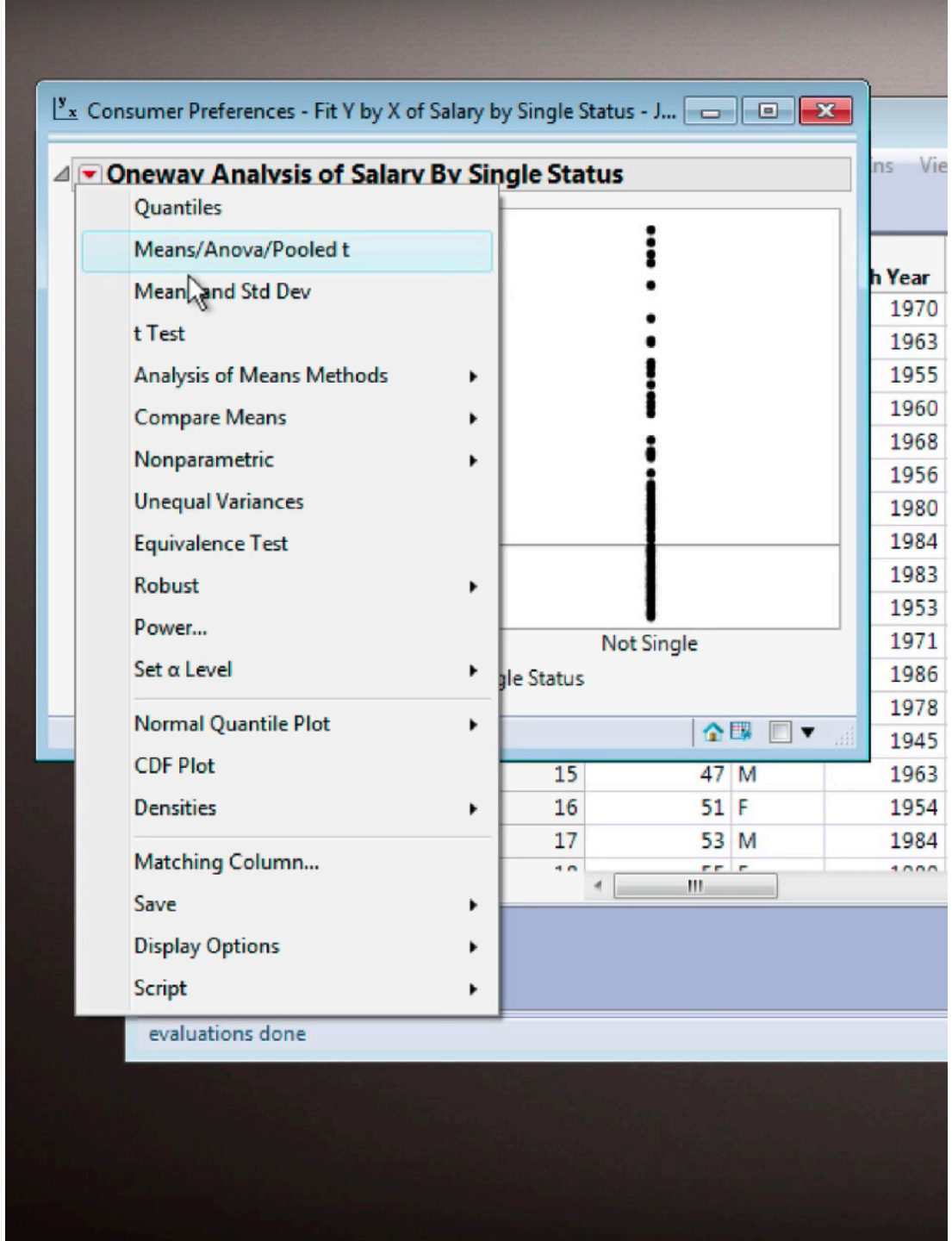
for the probabilities of the set of frequencies observed, and the two possible more extreme sets of frequencies which might have been observed. Without any assumption or approximation, therefore, the table observed may be judged significantly to contradict the hypothesis of proportionality if

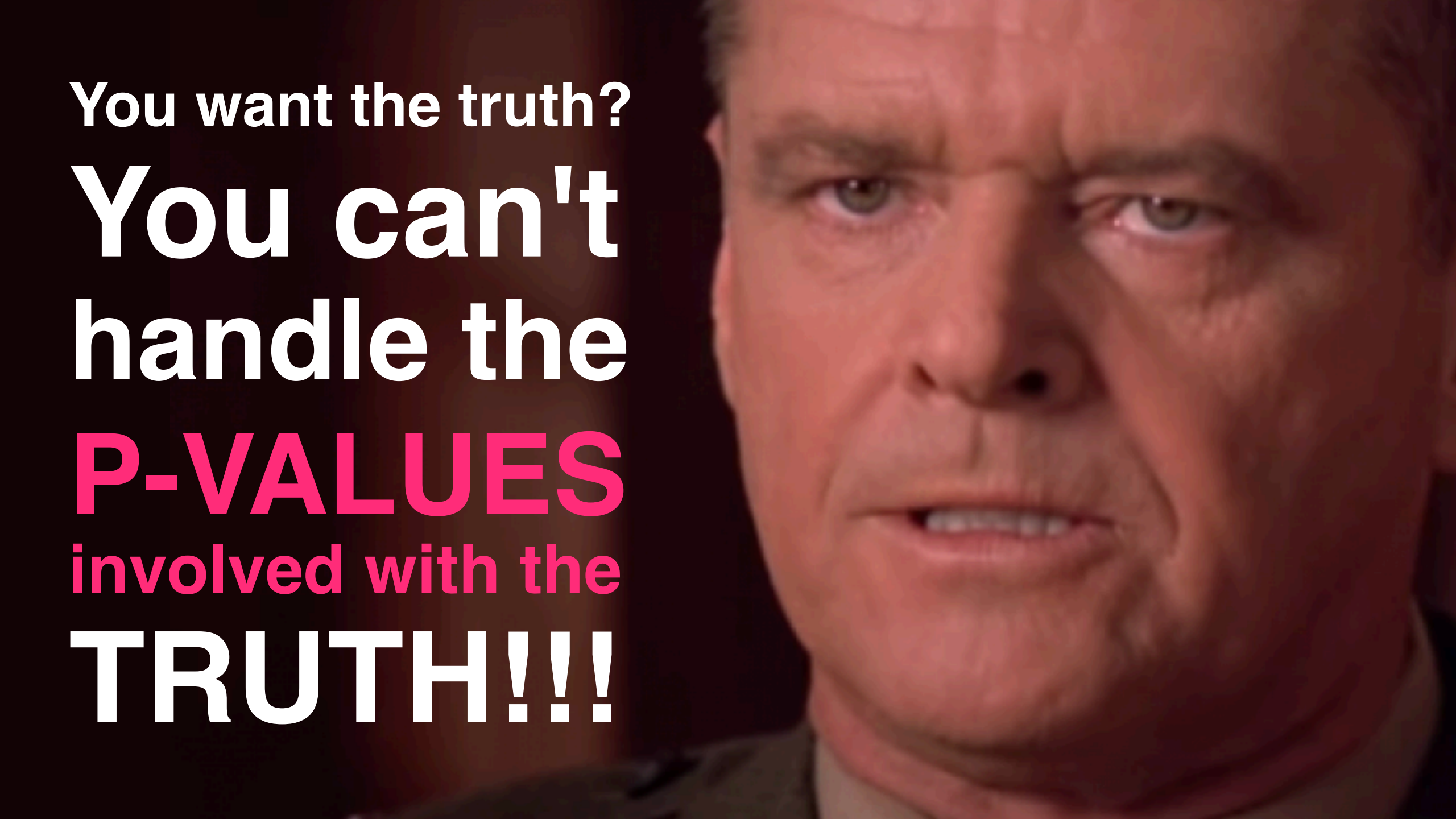
$$\frac{18!\,13!}{30!}(2992+102+1)$$

is a small quantity. This amounts to 619/1330665, or about 1 in 2150, showing that if the hypothesis of proportionality were true, observations of the kind recorded would be highly exceptional.

21·03. In its primary purpose of the comparison of a series of observed frequencies with those expected on the hypothesis to be tested, the χ^2 test is an approximate one, though validly applicable in an immense range of important cases. For other tests where the observations are measurements, instead of frequencies, it provides exact tests of significance. Of these the two most important are:—

(i.) its use to test whether a sample from a normal distribution confirms or contradicts the variance which this distribution is expected on theoretical grounds to have, and




You want the truth?
You can't handle the
P-VALUES
involved with the
TRUTH!!!

Accept uncertainty. Be

Thoughtful,

Open and

Modest.

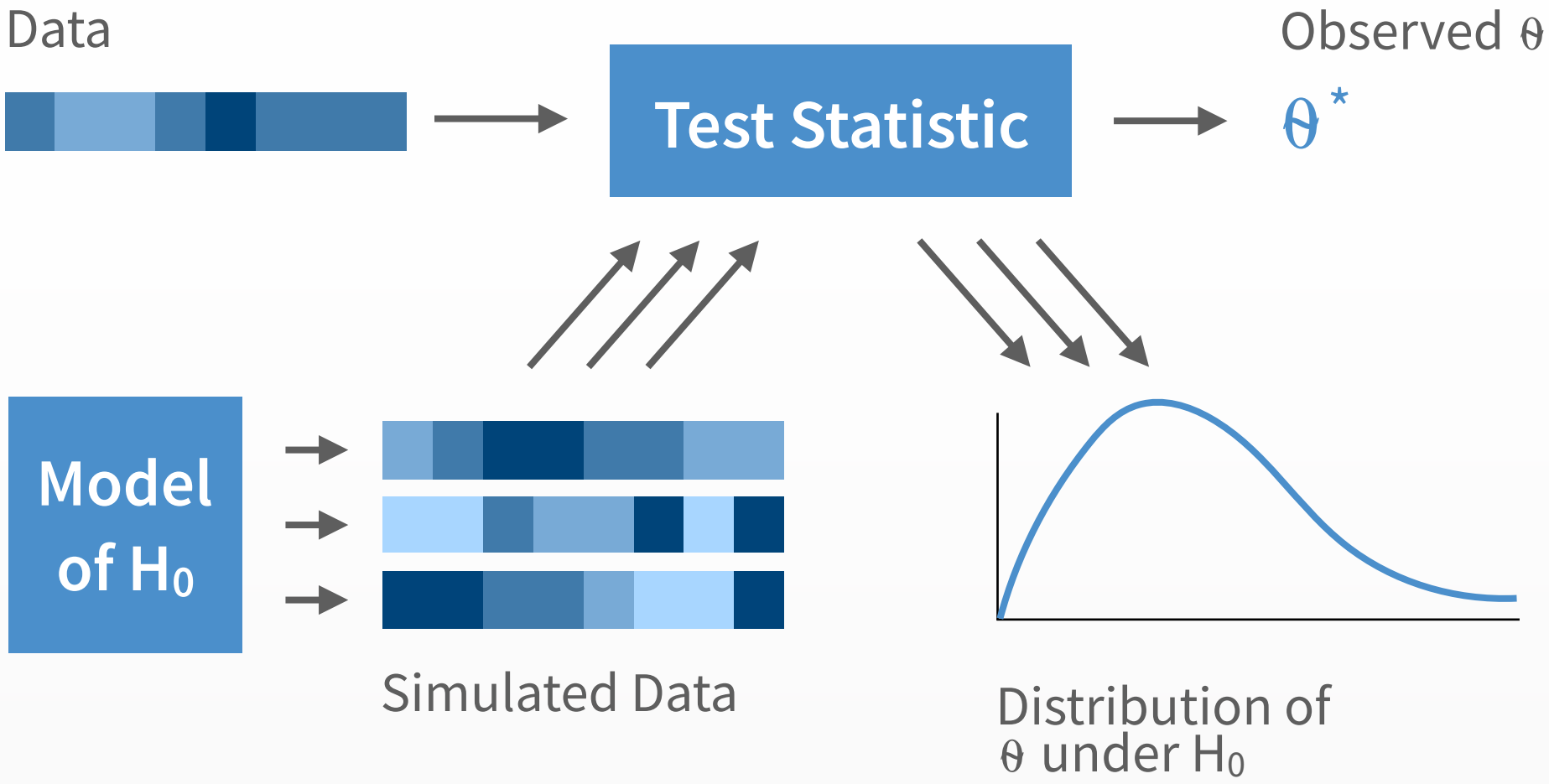
Plus 80 suggestions from 28 authors

Hypothesis Tests

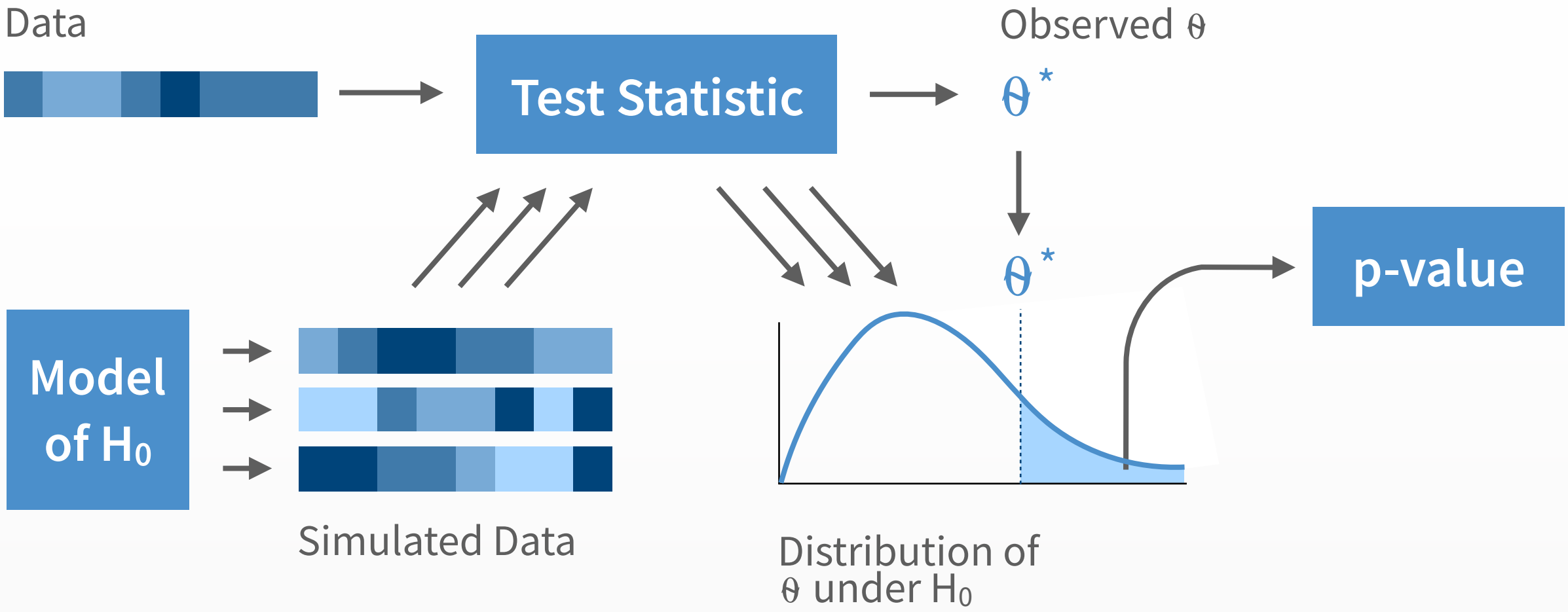
Downey, A. (2016) There is Still Only One Test.

Model of H_0

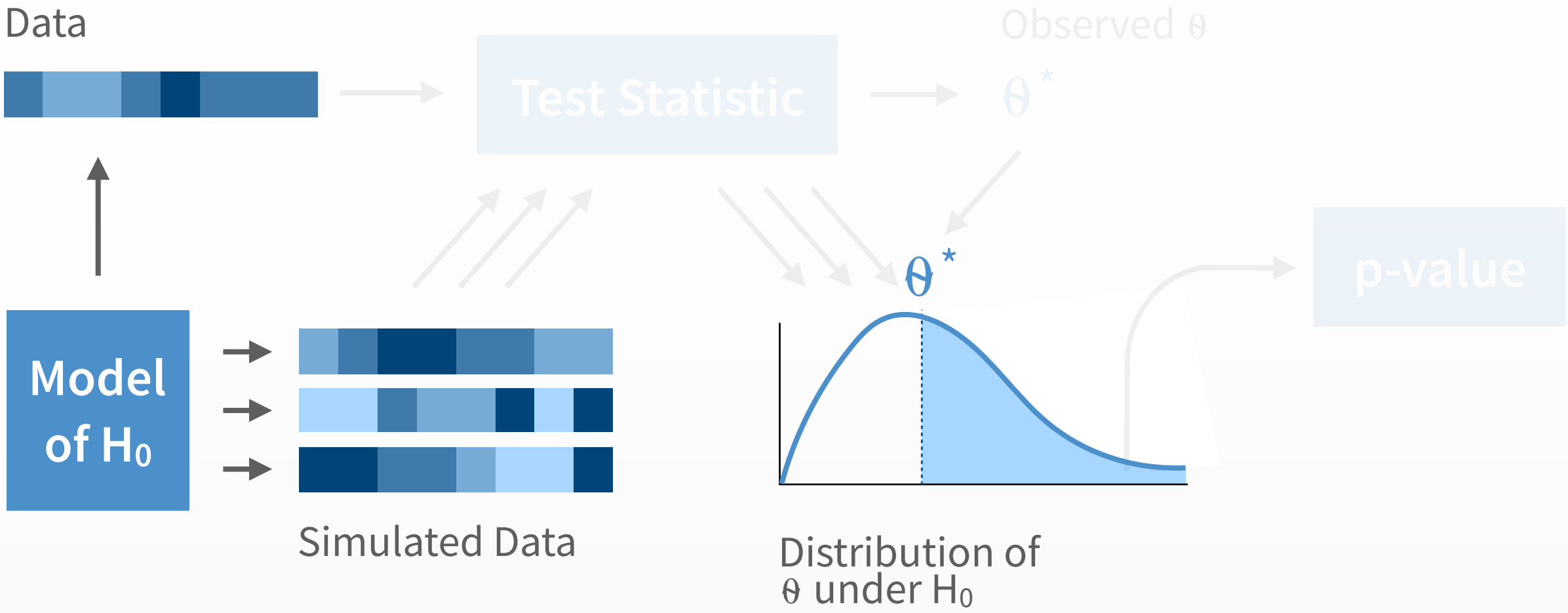
Downey, A. (2016) There is Still Only One Test.



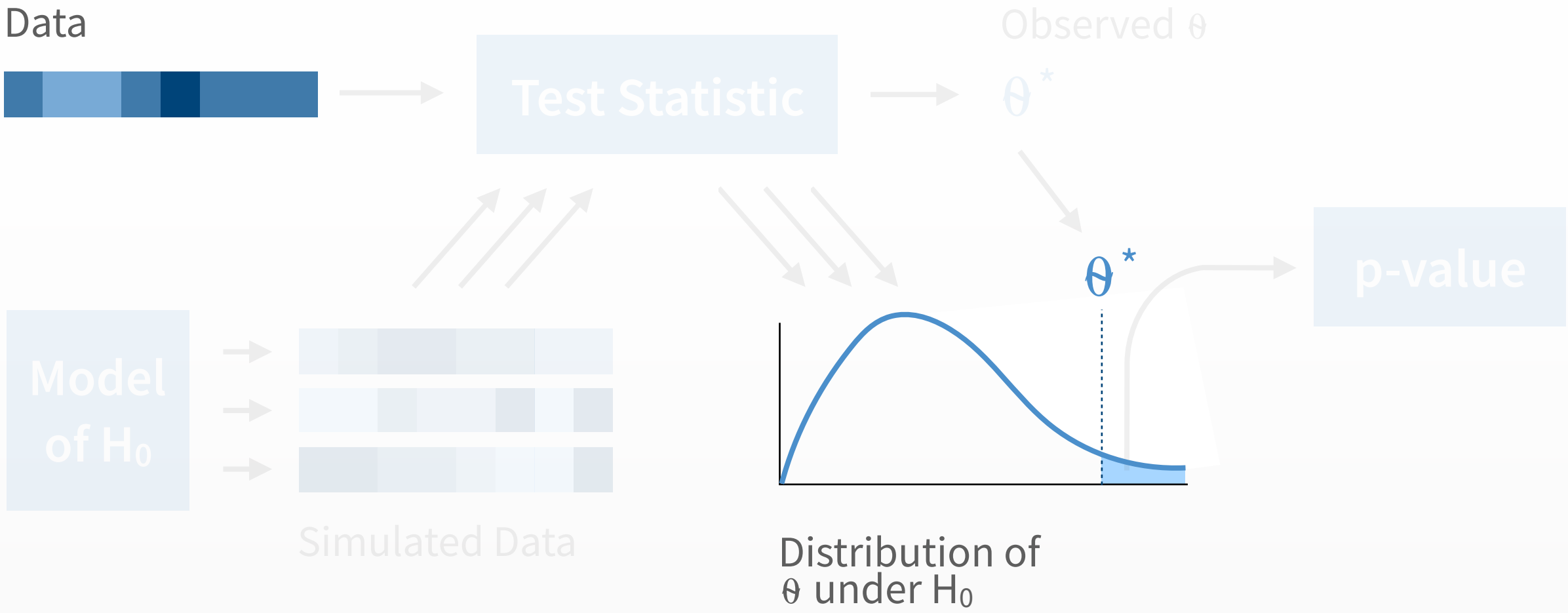
Downey, A. (2016) There is Still Only One Test.



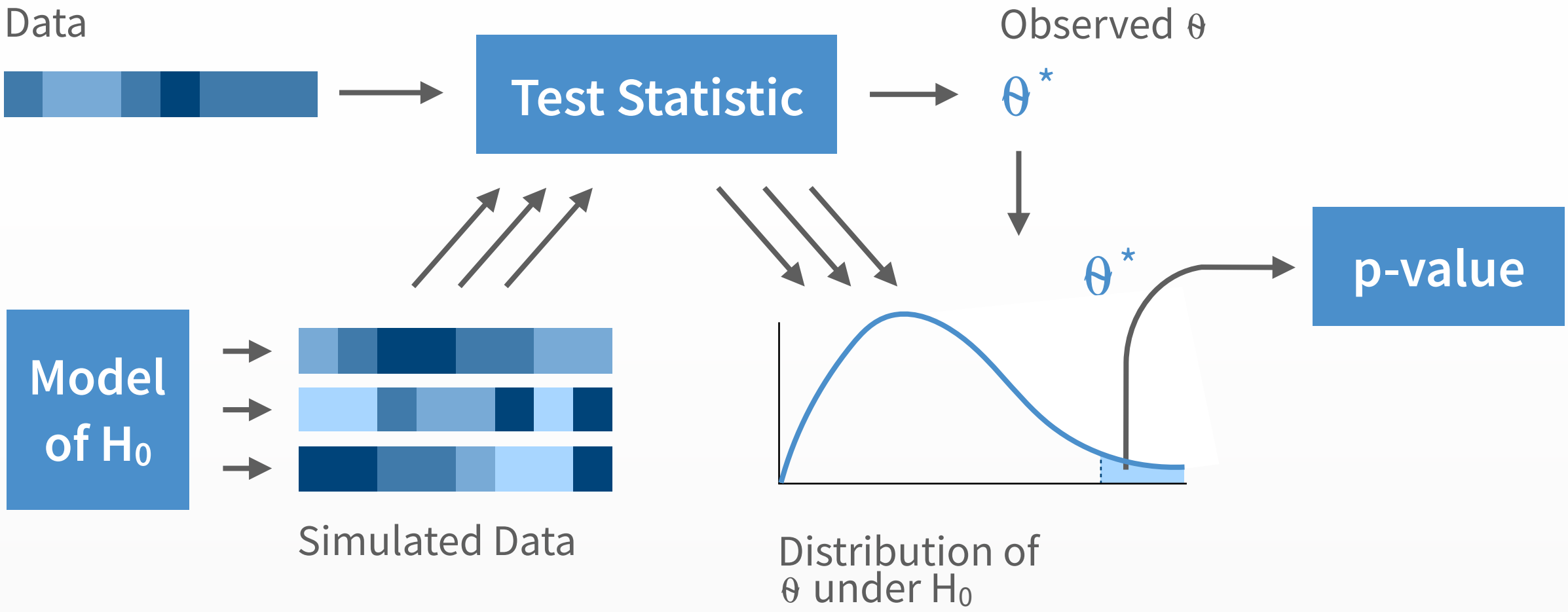
Downey, A. (2016) There is Still Only One Test.



Downey, A. (2016) There is Still Only One Test.



Downey, A. (2016) There is Still Only One Test.



2019

1620

Francis Bacon

Science = observation + induction

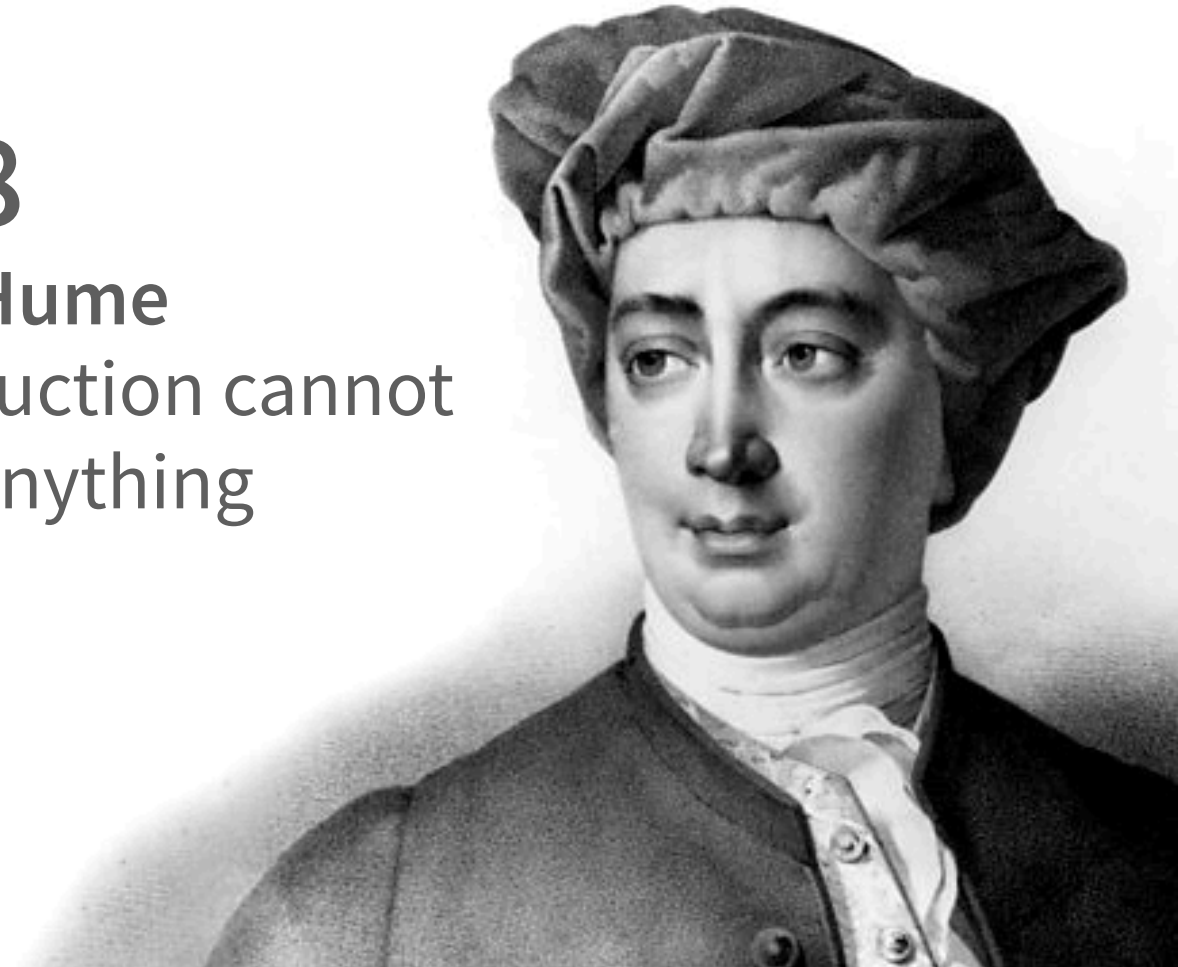
1620

2019

1748

David Hume

But induction cannot prove anything

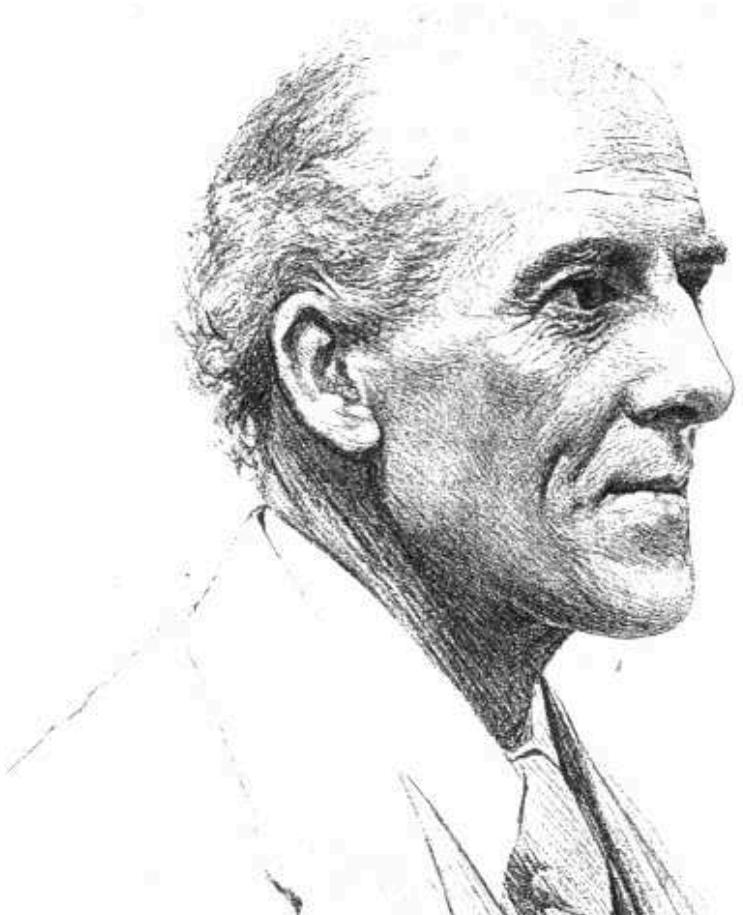


1620 1748 2019

1800's?

If you can't prove, disprove.

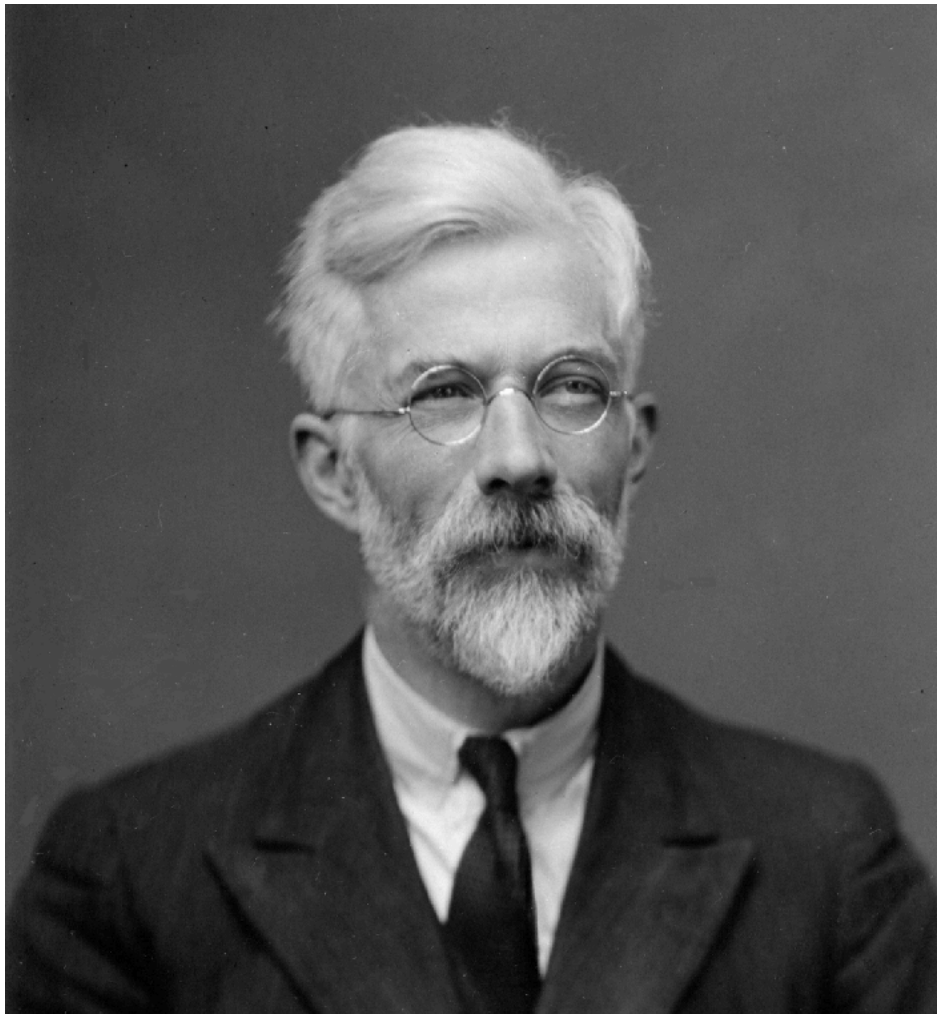
1620 1748 2019



1900's

Probabilities imply that you can not disprove either.

1620 1748 2019



1925

Ronald Fisher

If you cannot disprove,
make a reasonable guess

1620

1687
Isaac Newton
Classical Mechanics

1748

1796
Edward Jenner
Vaccinations

1824
Sadie Carnot
Thermodynamics

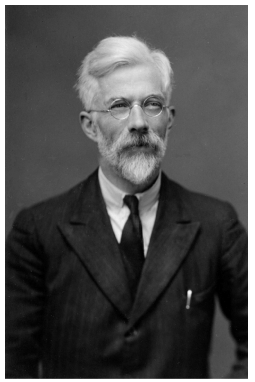
1862
James Maxwell
Electrodynamics

1864
Louis Pasteur
Germ Theory

1905
Albert Einstein
Special Relativity

1915
Albert Einstein
General Relativity

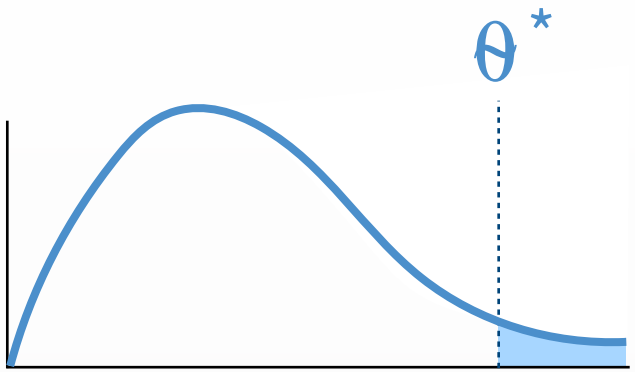
1925
Ronald Fisher



p-values provide a way to disprove in the face of uncertainty

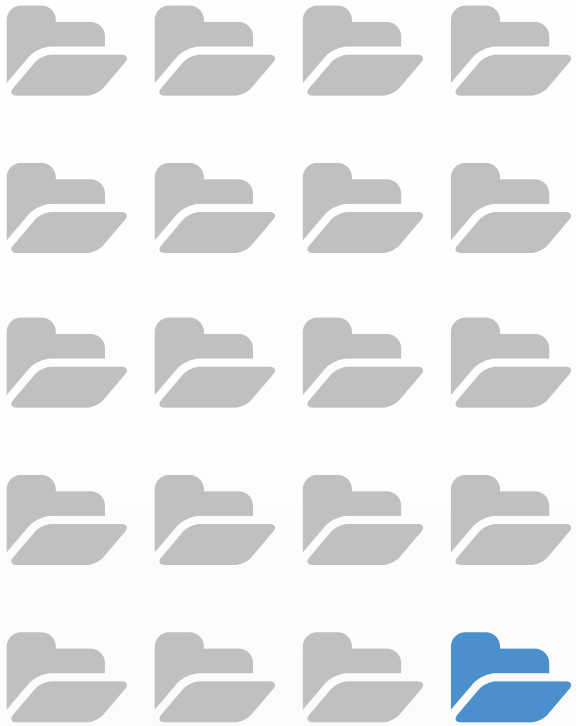
2019

Where do you draw the line?



Distribution of
θ under H_0

The file drawer problem



Conducted

Published

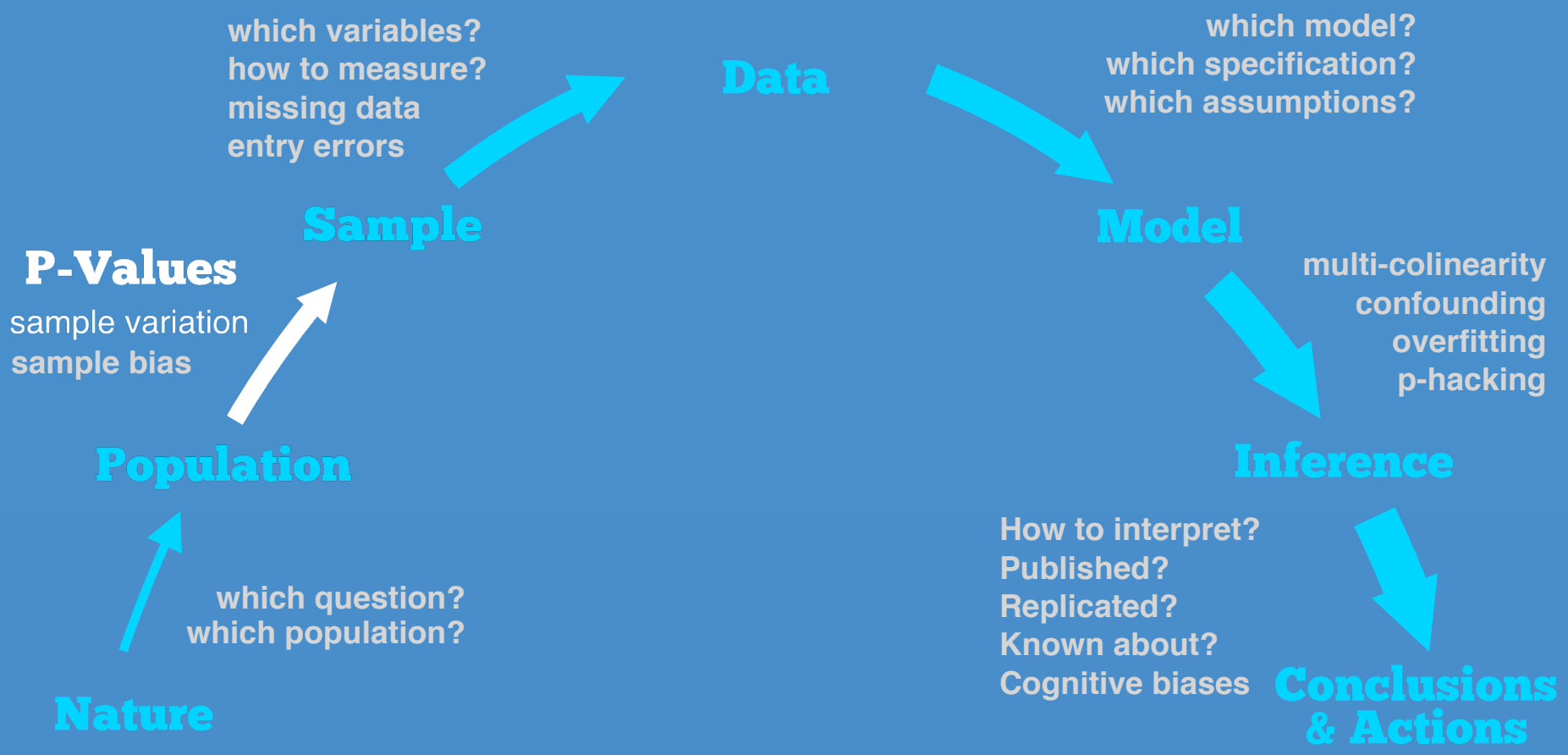
Bias

P-Values

sample variation

Sample

Population



Adapted from Levy DG, Norris D. Methods and Guidelines for Integrity in Multivariate Analysis of Real World (Observational) Data. Unpublished Manuscript.

where $q = 1 - p$. The probability that in a sample of $(c+d)$ trials it will occur c times is

$$\frac{(c+d)!}{c!\,d!} p^c q^d.$$

So that the probability of the observed frequencies a, b, c, and d in a 2×2 table is the product

$$\frac{(a+b)!\,(c+d)!}{a!\,b!\,c!\,d!} p^{a+c} q^{b+d},$$

and this in general must be unknown if p is unknown. The unknown factor involving p and q will, however, be the same for all tables having the same marginal frequencies $a+c$, $b+d$, $a+b$, $c+d$, so that among possible sets of observations having the same marginal frequencies, the probabilities are in proportion to

$$\frac{1}{a!\,b!\,c!\,d!},$$

whatever may be the value of p, or, in other words, for all populations in which the four frequencies are in proportion.

Now the sum of the quantities $1/a!\,b!\,c!\,d!$ for all samples having the same margins is found to be

$$\frac{n!}{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}$$

where $n = a+b+c+d$; so that, given the marginal frequencies, the probability of any observed set of entries is

$$\frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!} \cdot \frac{1}{a!\,b!\,c!\,d!}$$

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver and Boyd.

where $q = 1 - p$. The probability that in a sample of $(c+d)$ trials it will occur c times is

$$\frac{(c+d)!}{c!\,d!}p^c q^d.$$

So that the probability of the observed frequencies a, b, c, and d in a 2×2 table is the product

$$\frac{(a+b)!\,(c+d)!}{a!\,b!\,c!\,d!}p^{a+c}q^{b+d},$$

and this in general must be unknown if p is unknown. The unknown factor involving p and q will, however, be the same for all tables having the same marginal frequencies $a+c$, $b+d$, $a+b$, $c+d$, so that among possible sets of observations having the same marginal frequencies, the probabilities are in proportion to

$$\frac{1}{a!\,b!\,c!\,d!},$$

whatever may be the value of p, or, in other words, for all populations in which the four frequencies are in proportion.

Now the sum of the quantities $1/a!\,b!\,c!\,d!$ for all samples having the same margins is found to be

$$\frac{n!}{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}$$

where $n = a+b+c+d$; so that, given the marginal frequencies, the probability of any observed set of entries is

$$\frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!}\cdot\frac{1}{a!\,b!\,c!\,d!}$$

Which words do you associate with **math**?

hypotheses	**axioms**
messy	**logical**
best guess	**certain**
discover	**prove**

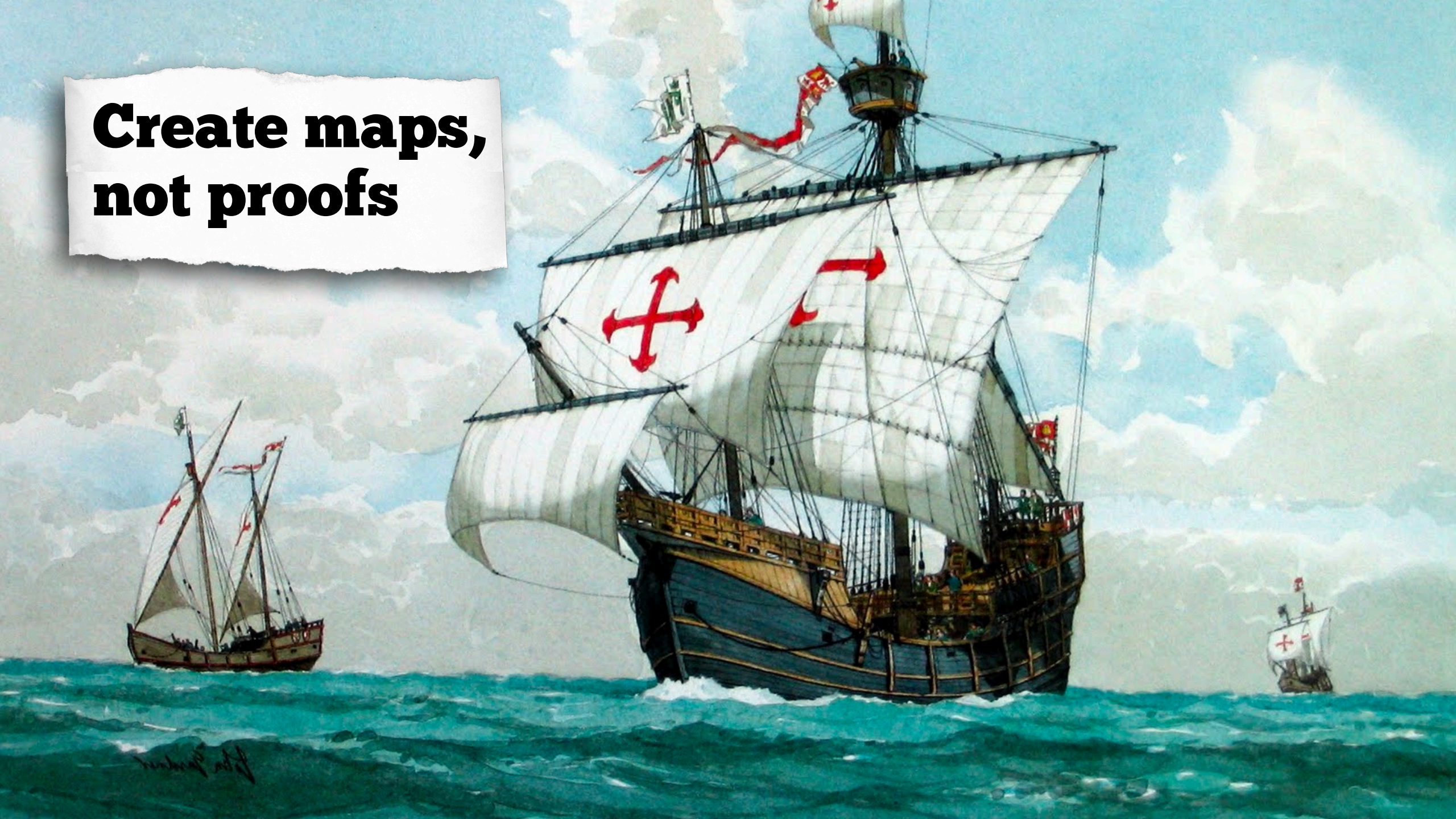
Which words do you associate with **Science**?

hypotheses	**axioms**
messy	**logical**
best guess	**certain**
discover	**prove**

We created a cargo cult
by confusing science
with math.

Now we must undo it.

Create maps, not proofs



Pop Quiz

What does **IMRAD** refer to? Poll your neighbors.

Pop Quiz

What does **IMRAD** refer to? Poll your neighbors.

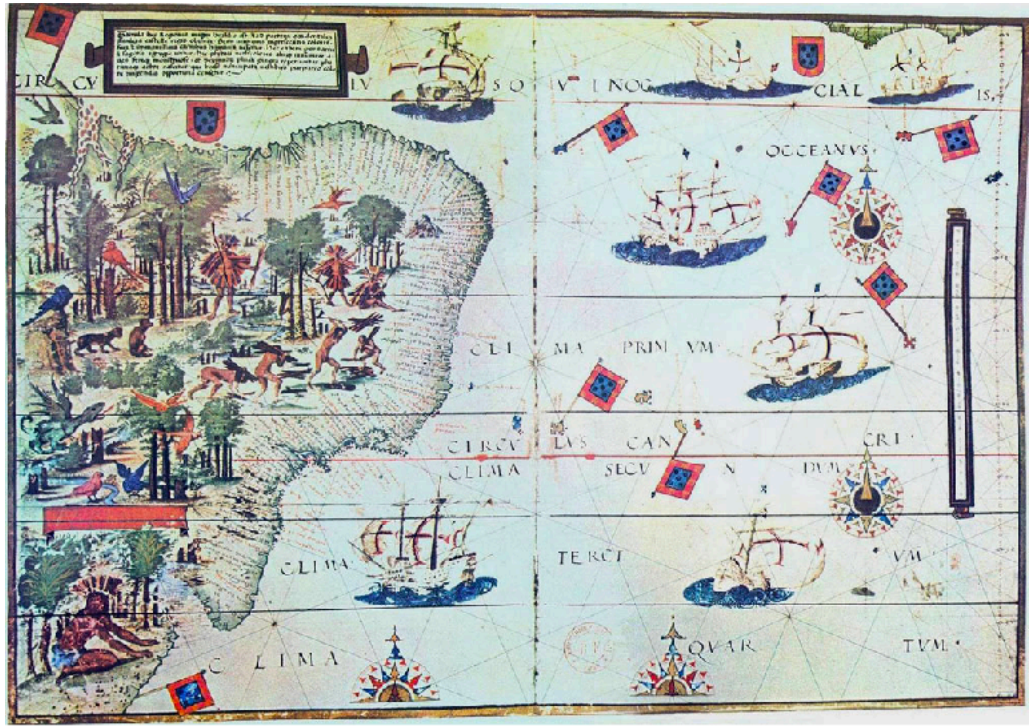
Introduction
What hypothesis was tested and why?

Methods
How was the study done?

Results
What answer was discovered?

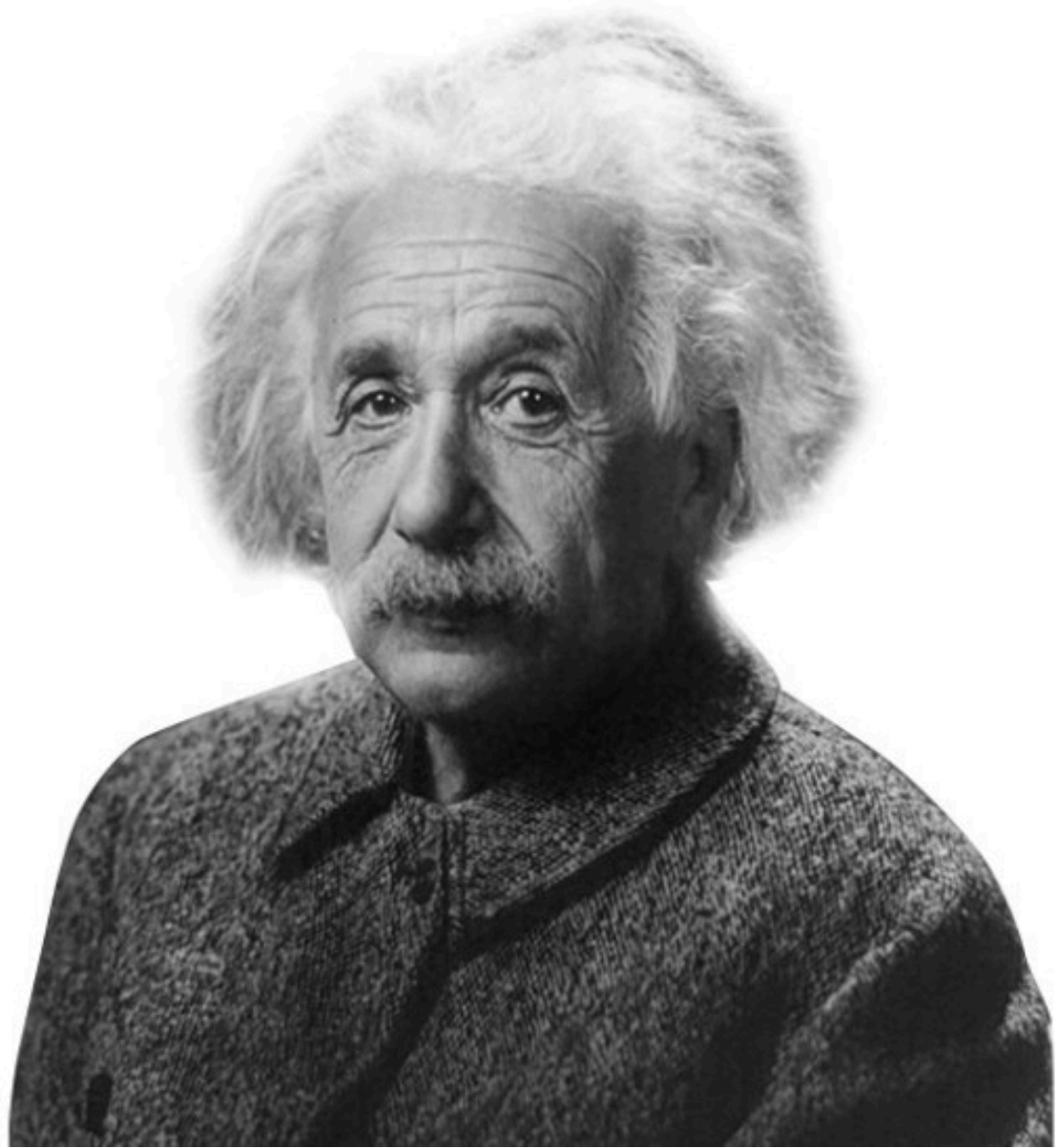
And Discussion
What does the answer imply?

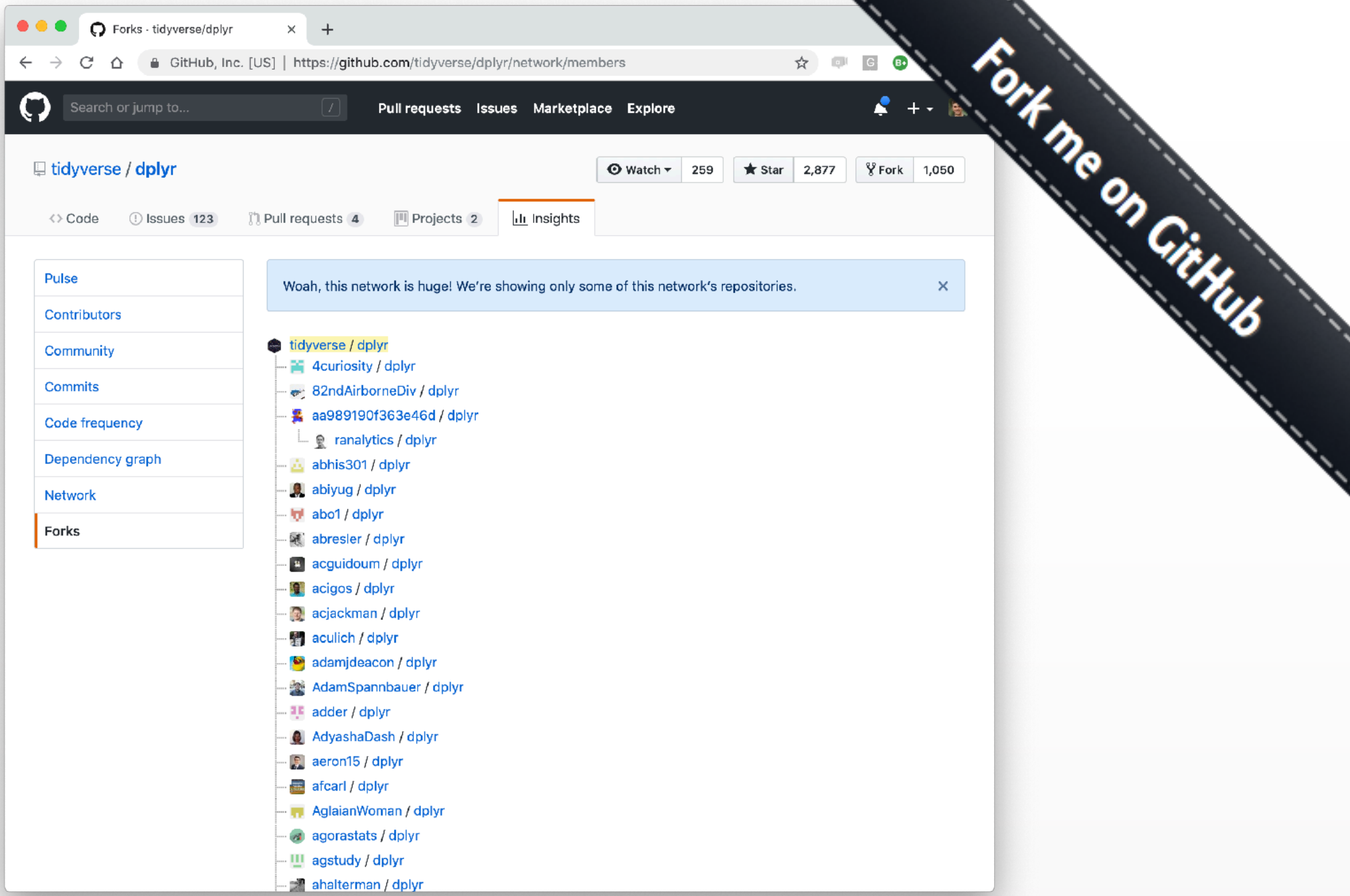
Maps yield models



Maps yield models

$$E = mc^2$$





Think
Open Source

1. Selection
2. Mutation

26%

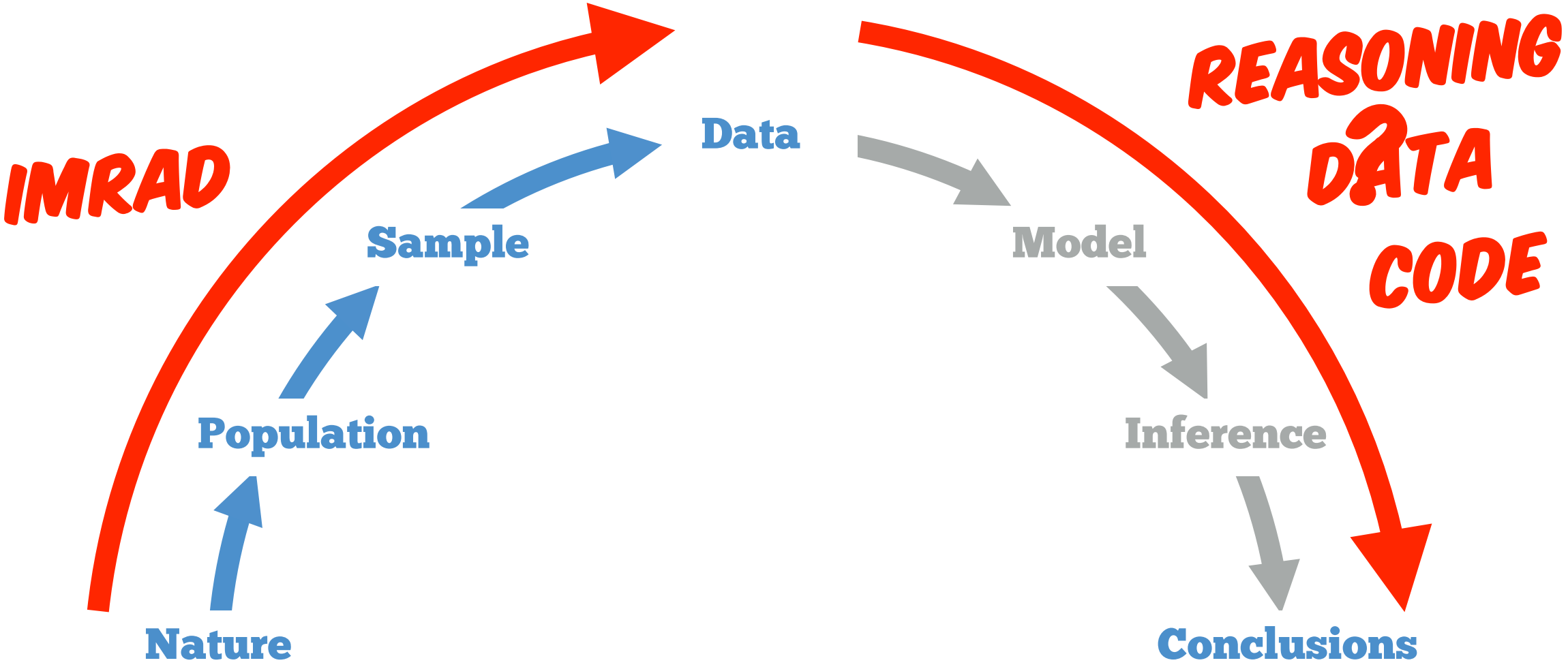
of authors provided code and data upon request *when required to do so*

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584-2589.

"R is a free software package available at www.r-project.org/ I used R for the [redacted] models….I used Matlab for the geometry."

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584-2589.

What is the source?



Adapted from Levy DG, Norris D. Methods and Guidelines for Integrity in Multivariate Analysis of Real World (Observational) Data. Unpublished Manuscript.


R Markdown is an authoring format for data science


Mythos
PREMIUM QUALITY HELLENIC BEER
PREMIUM
QUALITY
Mythos
HELLENIC BEER
750mL?

Dublin - Google Sheets

Garrett

Secure https://docs.google.com/spreadsheets/d/1OrHXGZJlyE_xyrbYs4YBElF...

Dublin

File Edit View Insert Format Data

SHARE

100% $ % .0 .00 123

fx bottle

	A	B	C	D	E
1	bottle	amount			
2	1	748.7			
3	2	746.4			
4	3	750.1			
5	4	751.9			
6	5	747.6			
7	6	748.8			
8	7	748.6			
9	8	752.2			
10	9	749.9			
11	10	745.9			
12	11	747.1			
13	12	748.3			
14	13	750.4			
15	14	749.6			
16	15	746			
17	16	750.5			
18	17	750.5			
19		750.1			
20		743.9			
21		750.6			

R Markdown Demo 1

One more thing

Reproducibility is an opportunity

RStudio Connect

Garrett

Secure | https://connect.rstudioservices.com/connect/#/apps/149/access/105

Content / rmarkdown-demo

WG bill

Re: Bottle Machine

Does our bottle machine fill each bottle with 750 mL of beer (on average)?

Data

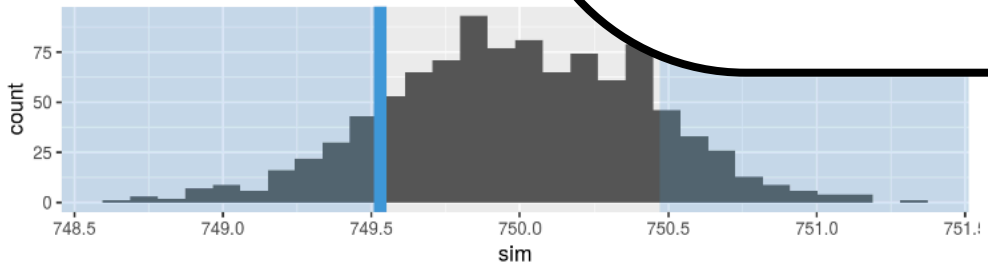
I selected 50 bottles at random from the Dublin factory, which contained the following amounts of beer (ml):

```
##  [1] 748.7 746.4 750.1 751.9 747.6 748.8 748.6 75
## [12] 748.3 750.4 749.6 746.0 750.5 750.5 750.1
## [23] 752.5 751.6 748.1 751.8 744.6 751.0 751
## [34] 748.6 744.7 750.6 748.9 753.2 747.8 74
## [45] 749.4 754.4 750.7 750.3 745.9 747.1
```

The mean amount is **749.53**.

Reasoning

The amounts of beer in our bottles should be normally distr
simulation to calculate the sample means of 1000 samples o



I've plotted the distribution of the simulated means above. The blue line shows our observed sample mean. Notice that 27.4% of
simulations produced a sample mean as extreme as ours (p-value = 0.274). In other words, our observations are in accordance wi
our belief.

Conclusion

We do **not** have evidence that the bottle machine is malfunctioning.

Info | Access | Runtime | Schedule | Tags | Vars | Logs

Who can view this document

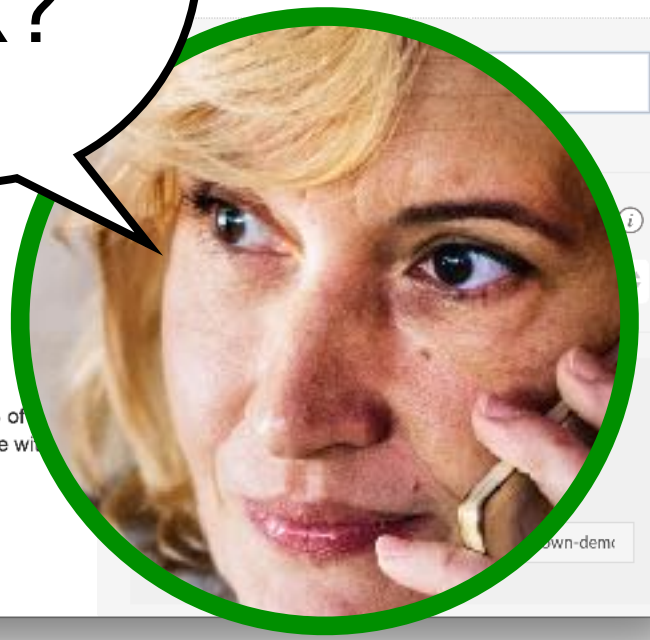
Collaborators & you

can change this document

William Gosset bill

VP Virginia virginia

Can you do this ***every*** week?



Schedule saved successfully.

Info Access Runtime Schedule Tags Vars Logs

Schedule output for default

Start date & time

Fri Sep 07 2018 12:39:20 GMT-0500

Set to Now

Schedule type

Weekly

Run every 1 week.

Run every...

- Sunday
- Monday
- Tuesday
- Wednesday
- Thursday
- Friday
- Saturday

Publish output after it is generated

Send email after update

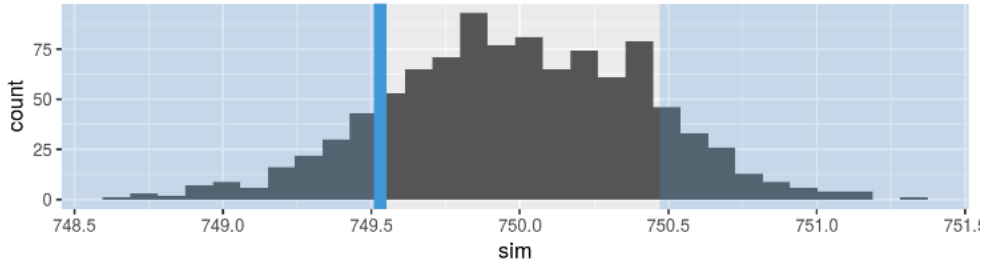
ed the following amounts of beer (mL):

```
##  [1] 752.2 749.9 745.9 747.1
## [13] 743.9 750.6 758.0 746.3
## [23]  753.7 749.4 752.1 747.9
## [34] 746.0 748.4 754.2 750.6 749.9
## [45] 749.
```

The mean amount is **749.53**.

Reasoning

The amounts of beer in our bottles should be normally distributed with a mean of 750 mL and a standard deviation of 3 mL. Let's use simulation to calculate the sample means of 1000 samples of size 50 generated from such a distribution.



I've plotted the distribution of the simulated means above. The blue line shows our observed sample mean. Notice that 27.4% of the simulations produced a sample mean as extreme as ours (p-value = 0.274). In other words, our observations are in accordance with our belief.

Conclusion

We do **not** have evidence that the bottle machine is malfunctioning.

RStudio Connect

Garrett

Secure | https://connect.rstudioservices.com/connect/#/apps/149/access/105

Content / rmarkdown-demo

WG bill

Re: Bottle Machine

Does our bottle machine fill each bottle with 750 mL of beer (on average)?

Data

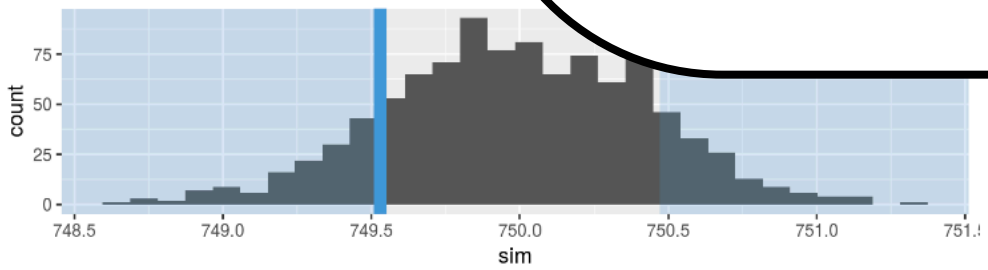
I selected 50 bottles at random from the Dublin factory, which c

```
##  [1] 748.7 746.4 750.1 751.9 747.6 748.8 7
## [12] 748.3 750.4 749.6 746.0 750.5 750.5
## [23] 752.5 751.6 748.1 751.8 744.6 751.
## [34] 748.6 744.7 750.6 748.9 753.2 747
## [45] 749.4 754.4 750.7 750.3 745.9 747
```

The mean amount is **749.53**.

Reasoning

The amounts of beer in our bottles should be normally
simulation to calculate the sample means of 1000 sample



I've plotted the distribution of the simulated means above. The blue line shows our observed sample mean. Notice that 27.4% of
simulations produced a sample mean as extreme as ours (p-value = 0.274). In other words, our observations are in accordance wi
our belief.

Conclusion

We do **not** have evidence that the bottle machine is malfunctioning.

Info | Access | Runtime | Schedule | Tags | Vars | Logs

Schedule output for default

Start date & time

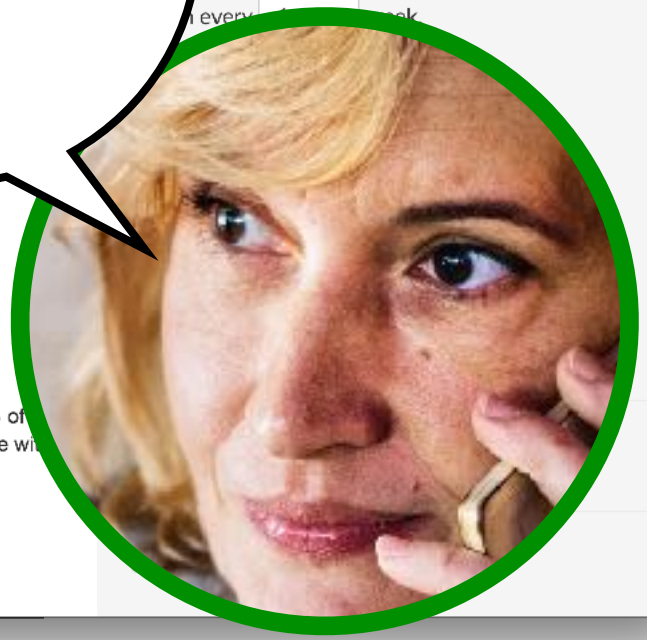
ri Sep 07 2018 12:39:20 GMT-0500

Set to Now

le type

ly

Can you do this for the ***London*** factory?



R Markdown Demo 2

RStudio Connect

Garrett

Secure https://connect.rstudioservices.com/connect/#/apps/150/access/106

Content / rmarkdown-demo

WG bill

Re: Bottle Mac

Does our bottle machine fill each bottle wi

Data

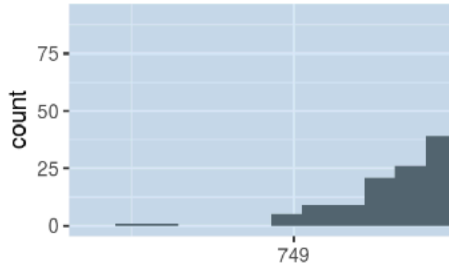
I selected 50 bottles at random from the D
(mL):

```
##  [1] 748.7 746.4 750.1 751.
## [12] 748.3 750.4 749.6 746.
## [23] 752.5 751.6 748.1 751.
## [34] 748.6 744.7 750.6 748.
## [45] 749.4 754.4 750.7 750.
```

The mean amount is **749.53**.

Reasoning

The amounts of beer in our bottles should
standard deviation of 3 mL. Let's use simu
50 generated from such a distribution.



I've plotted the distribution of the simulate
mean. Notice that 26% of the simulations

Re: Bottle Machine

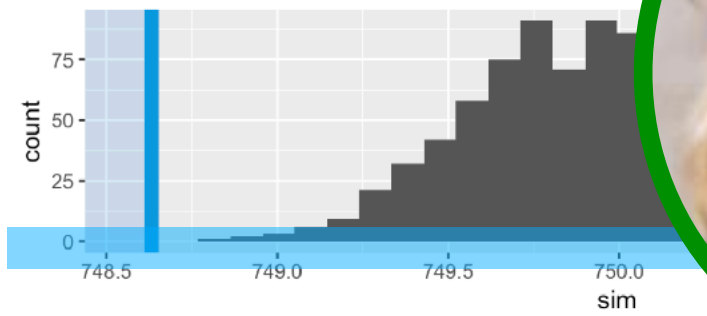
Does our bottle machine fill each bottle with 750 mL of beer (on average)?

Data

ined the following amounts of beer (mL):

```
7.7 747.4 751.6
.1 745.5 752.0
.5 750.7 748.9
2 751.4 749.5
```

The
of 3 mL. Let
distribution.



I've plotted the distribution of the simulated means above. The blue line show
mean. Notice that 0% of the simulations produced a sample mean as extreme as ours (p-value = 0). In other words, our

I like it!

Reproducibility is an opportunity

schedule

parameterize

automate

What should we do the leftover beer?

Photo by Wil Stewart on Unsplash

Thank You