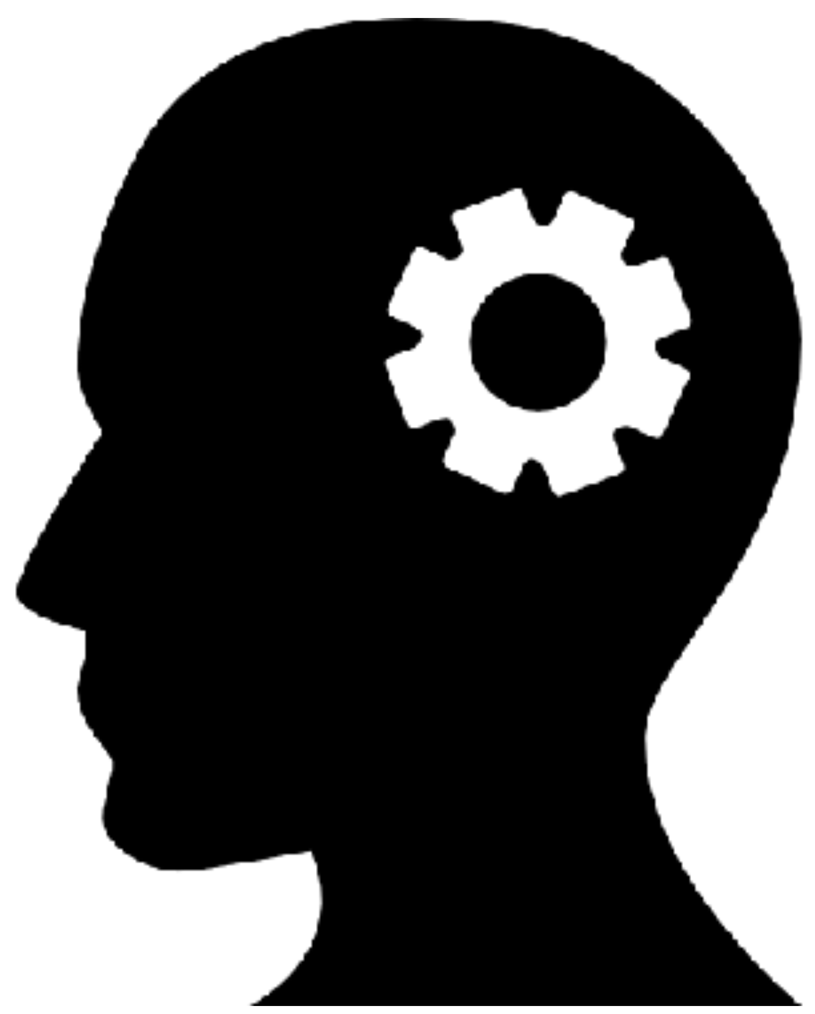
Statistical inference and things that can go wrong

Timo B. Roettger

@TimoRoettger

election
polls



human
cognition

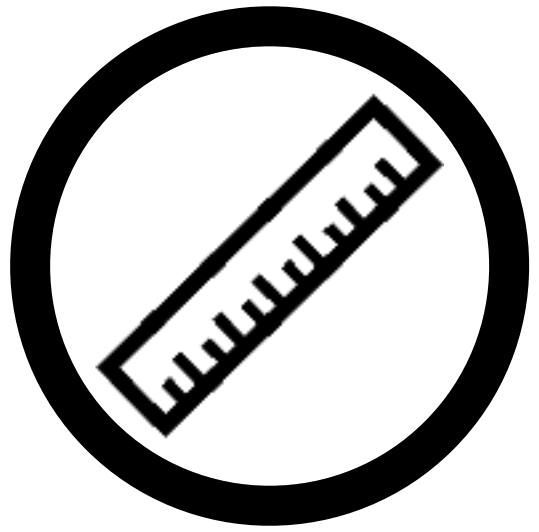
scientific publications

Is there a relationship between **autism** and **vaccination**?

Is there a relationship between **autism** and **vaccination**?

measure some stuff…

Is there a relationship between **autism** and **vaccination**?



measure some stuff…

Evaluate that measured stuff using **statistical inference**

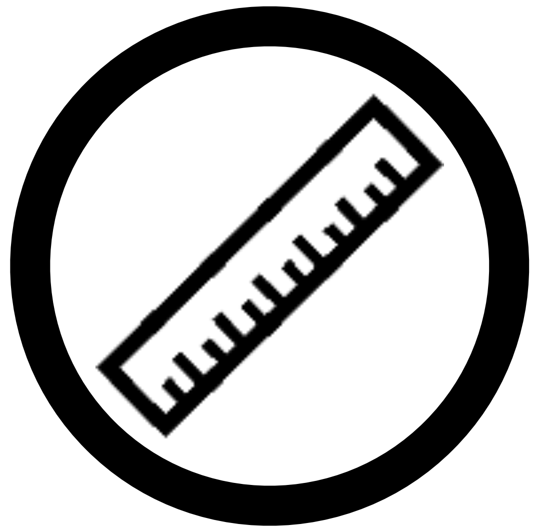
Is there a relationship between **autism** and **vaccination**?

measure some stuff…

Evaluate that measured stuff using **statistical inference**

Publish findings in a scientific journal

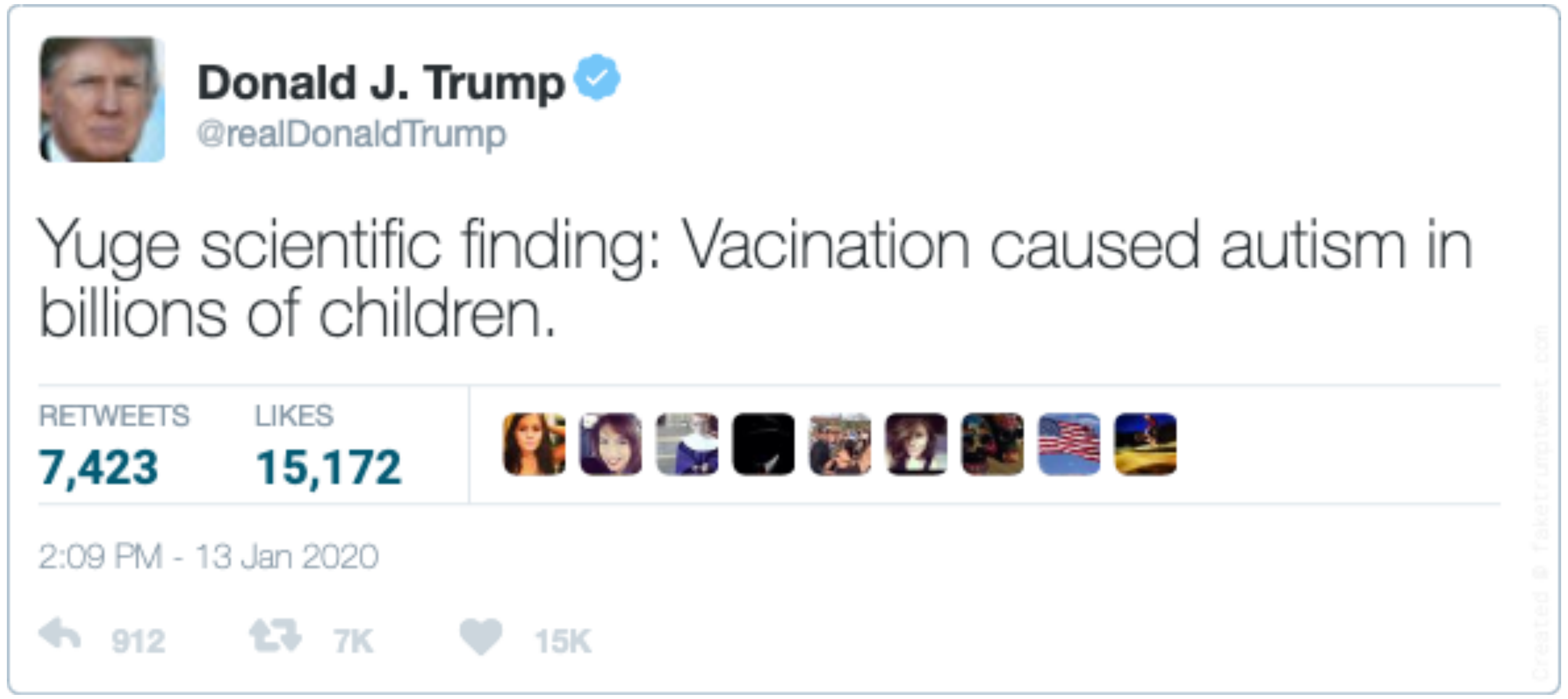
Is there a relationship between **autism** and **vaccination**?

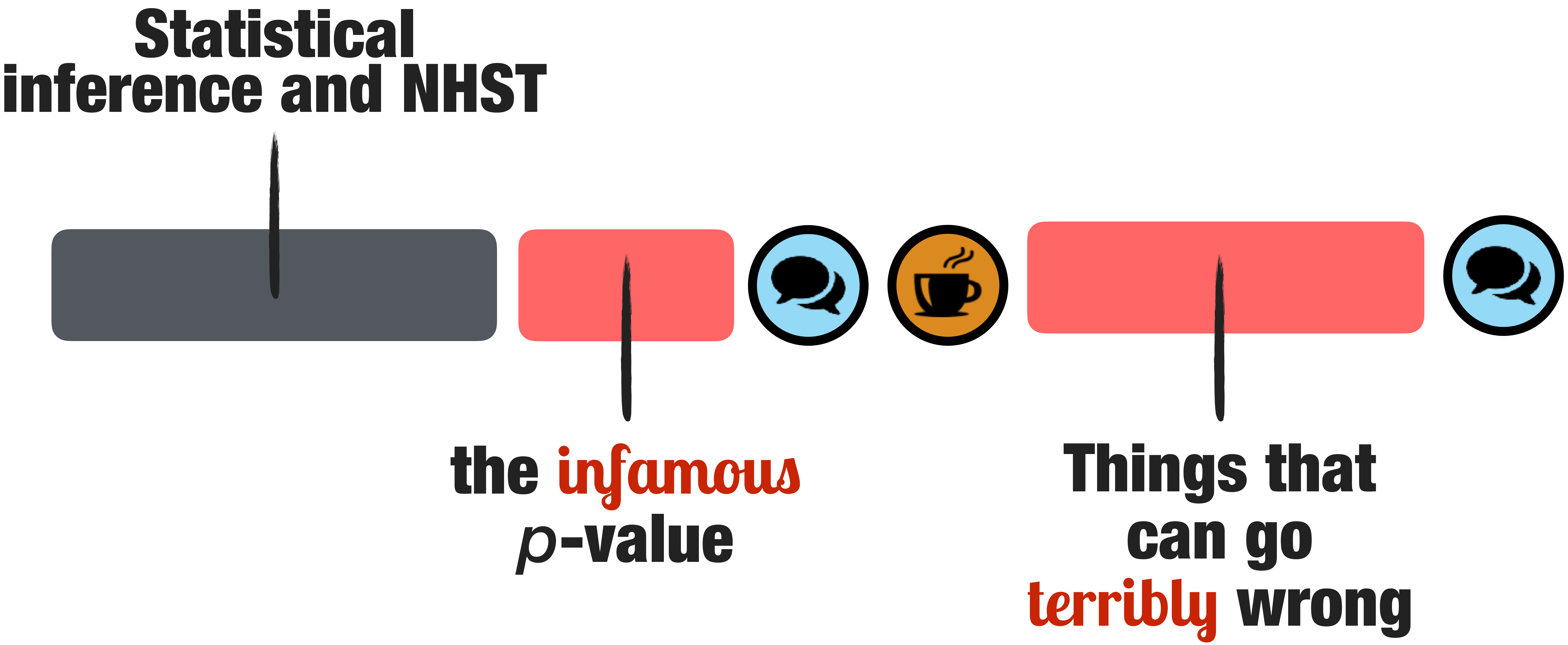


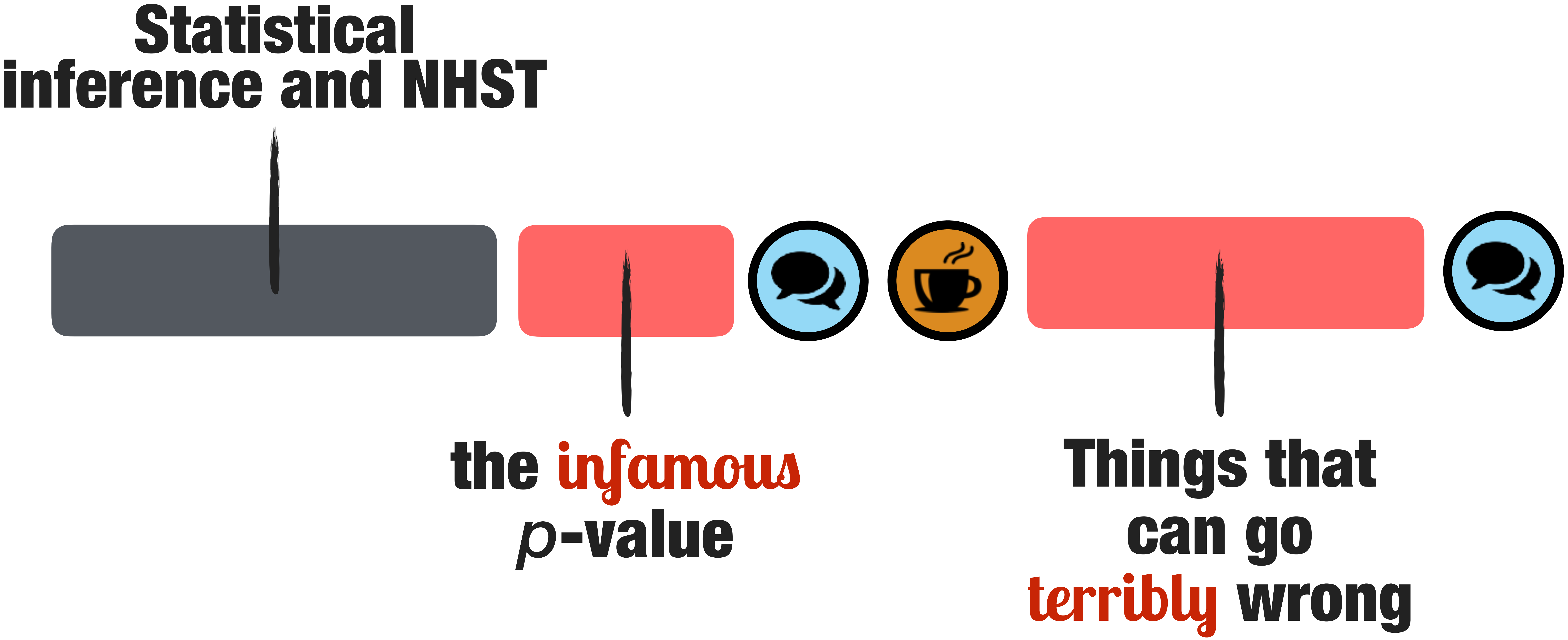
measure some stuff…

Evaluate that measured stuff using **statistical inference**

Publish findings in a scientific journal

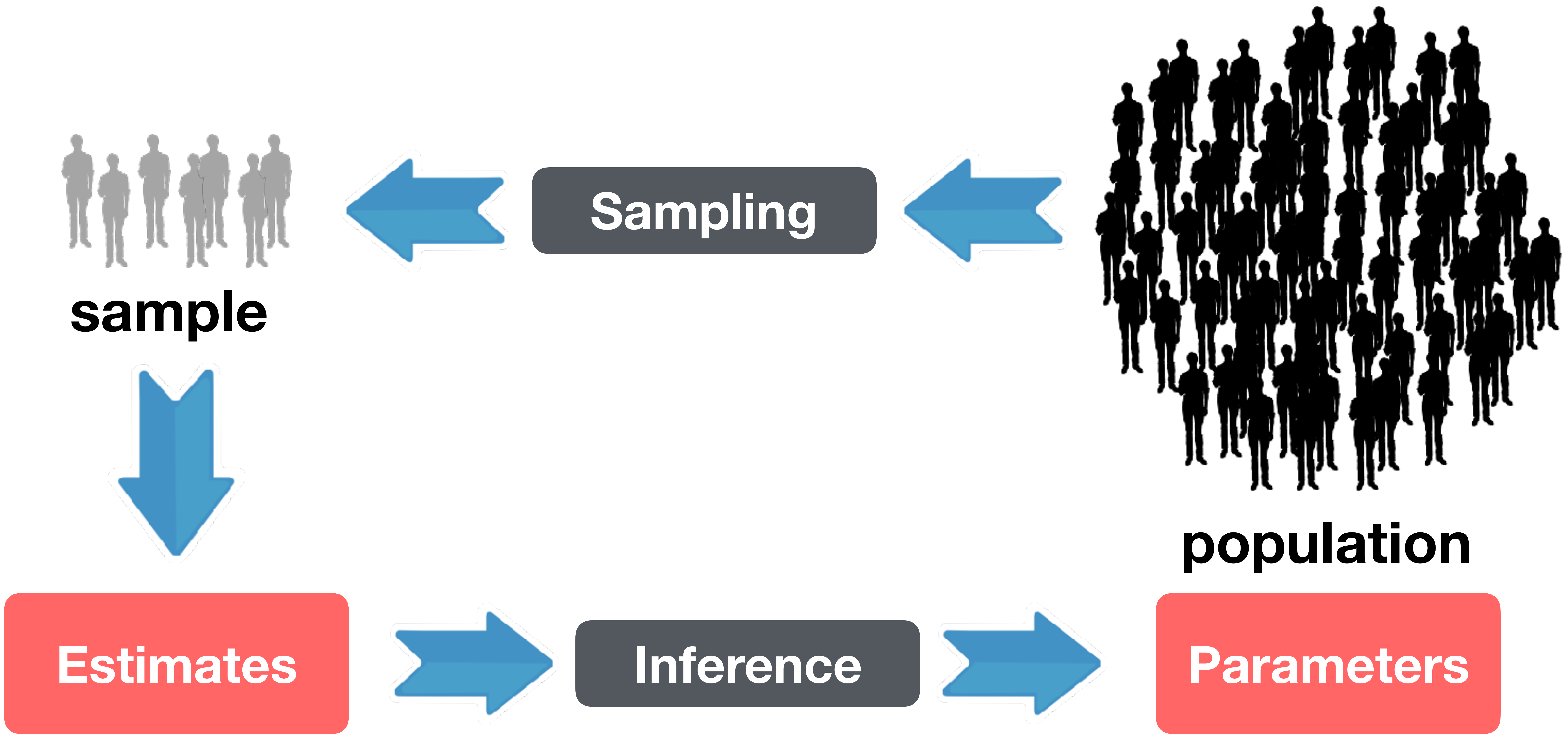



Statistical inference and NHST
the infamous p-value
Things that can go terribly wrong

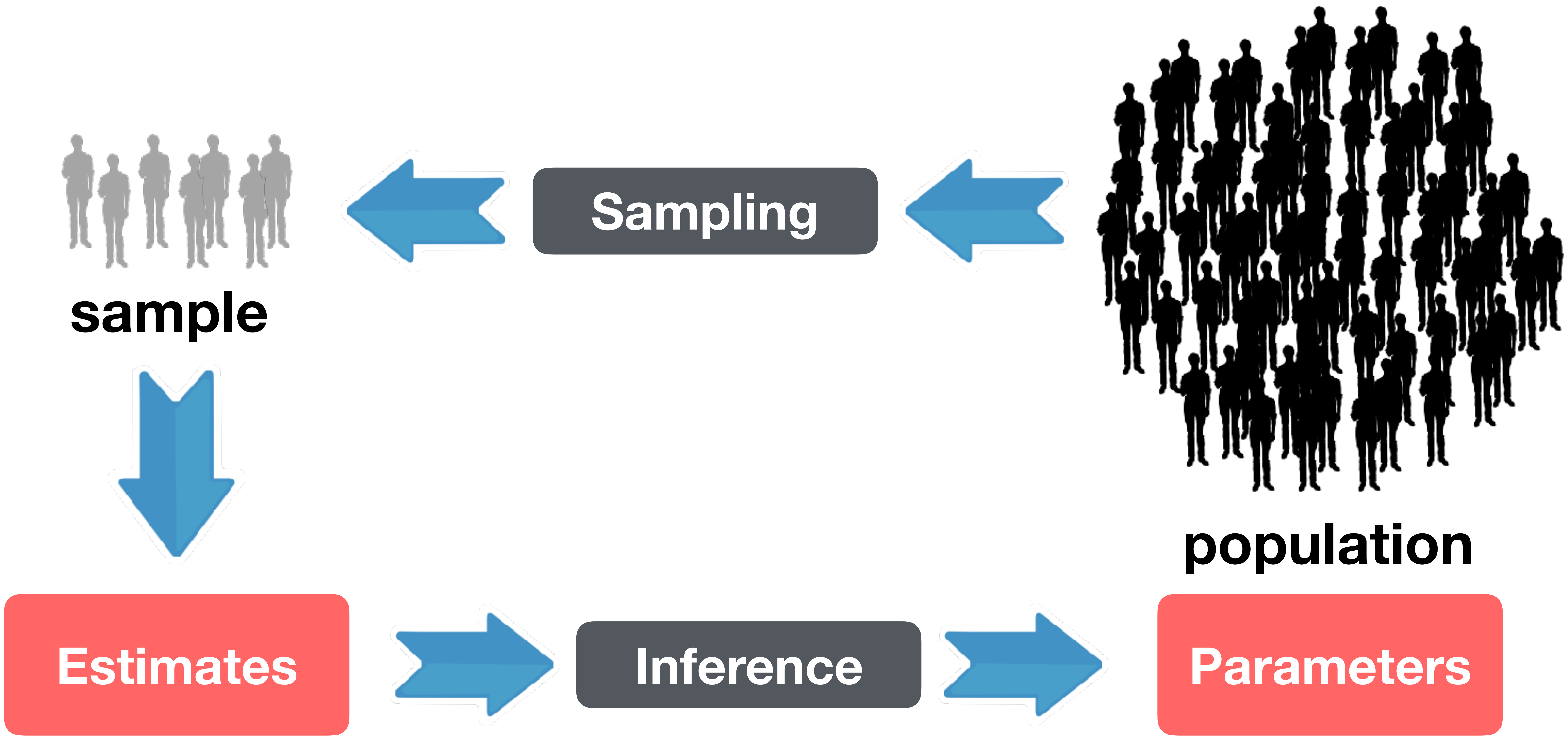

Statistical inference and NHST
the infamous p-value
Things that can go terribly wrong

Statistical inference

Statistical inference



Statistical inference



NHST

Null-**H**ypothesis **S**ignificance **T**esting

NHST

Null-Hypothesis Significance Testing

0. Set up the Alternative Hypothesis (H_a).

1. Set up the Null-Hypothesis (H_0).

NHST

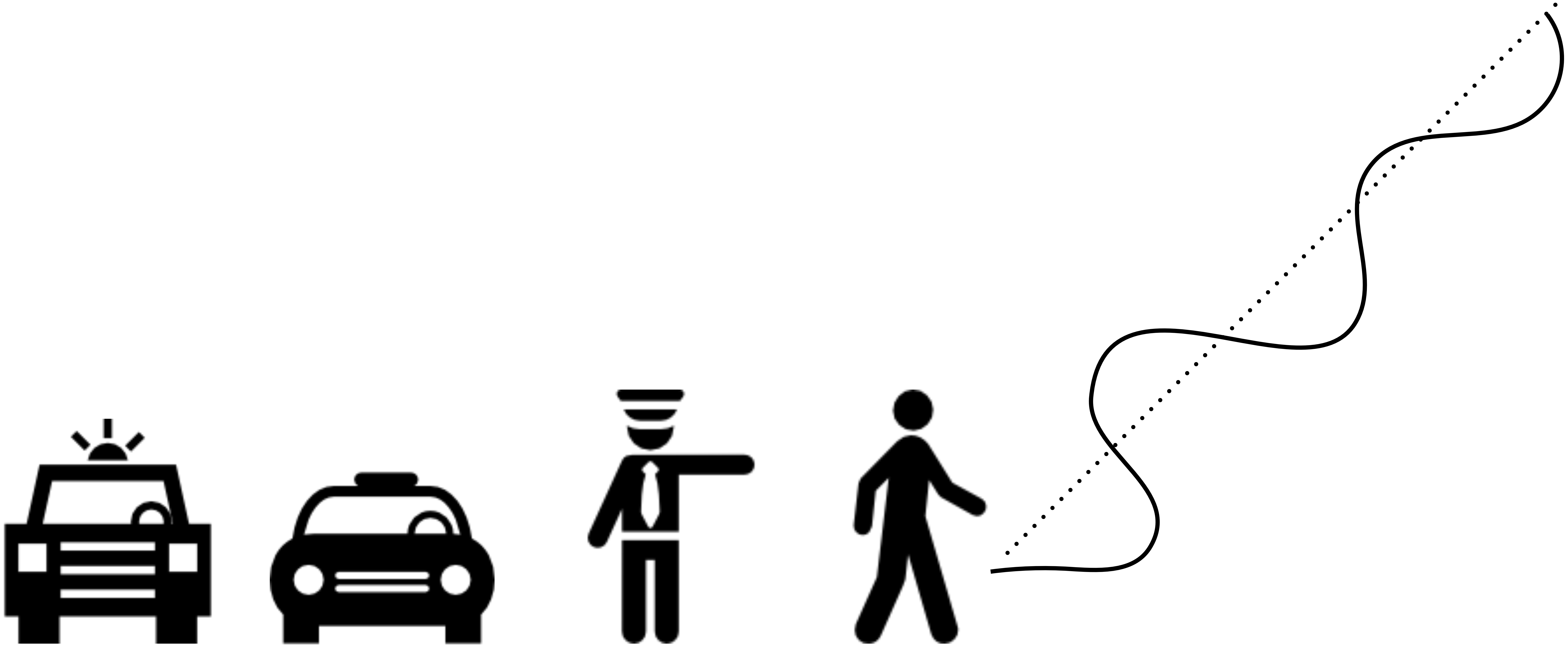
Null-Hypothesis Significance Testing

0. Set up the Alternative Hypothesis (H_a).

1. Set up the Null-Hypothesis (H_0).

2. Calculate the probability of the results under H_0 (p value).

3. Reject H_0 when $p < 0.05$, else do not reject.



H_a: Drunken people diverge more from a straight line than sober people.

H_a: drunk > sober

H_0: Drunken people diverge as much from a straight line as sober people.

H_0: drunk = sober

Measuring the compatibility of the data

* this is a simplified version of the formula to make a conceptual point, please do not use this to actually calculate *t*-values as the actual formulas are a bit more complicated

difference between groups

the larger *t*, the smaller *p*

$$\boxed{t} = \frac{\boxed{\bar{x}_1 - \bar{x}_2}}{\underbrace{\sqrt{\frac{\boxed{SD}}{\boxed{N}}}}_{\text{standard error (SE)}}}$$

"noise"

sample size

Measuring the compatibility of the data

difference between groups

the larger *t*,
the smaller *p*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\underbrace{\sqrt{\frac{SD}{N}}}_{\text{standard error (SE)}}}$$

standard error (SE)

"noise"

sample size

Measuring the compatibility of the data

$$\underbrace{9.8}_{t\ \text{value}} = \frac{\overbrace{4 - 2}^{\text{difference between groups}}}{\underbrace{\sqrt{\frac{1\ (\text{“noise”})}{24\ (\text{sample size})}}}_{\text{standard error (SE)}}}$$

Measuring the compatibility of the data

difference between groups

t value

$$2.5 = \frac{4 - 3.5}{\underbrace{\sqrt{\dfrac{1}{24}}}_{\text{standard error (SE)}}}$$

“noise”

sample size

Measuring the compatibility of the data

difference between groups

t value

$$\overset{t\text{ value}}{5.5} = \frac{\overbrace{4 - 3.5}^{\text{difference between groups}}}{\underbrace{\sqrt{\frac{0.2}{24}}}_{\text{standard error (SE)}}}$$

standard error (SE)

“noise”

sample size

Measuring the compatibility of the data

difference between groups

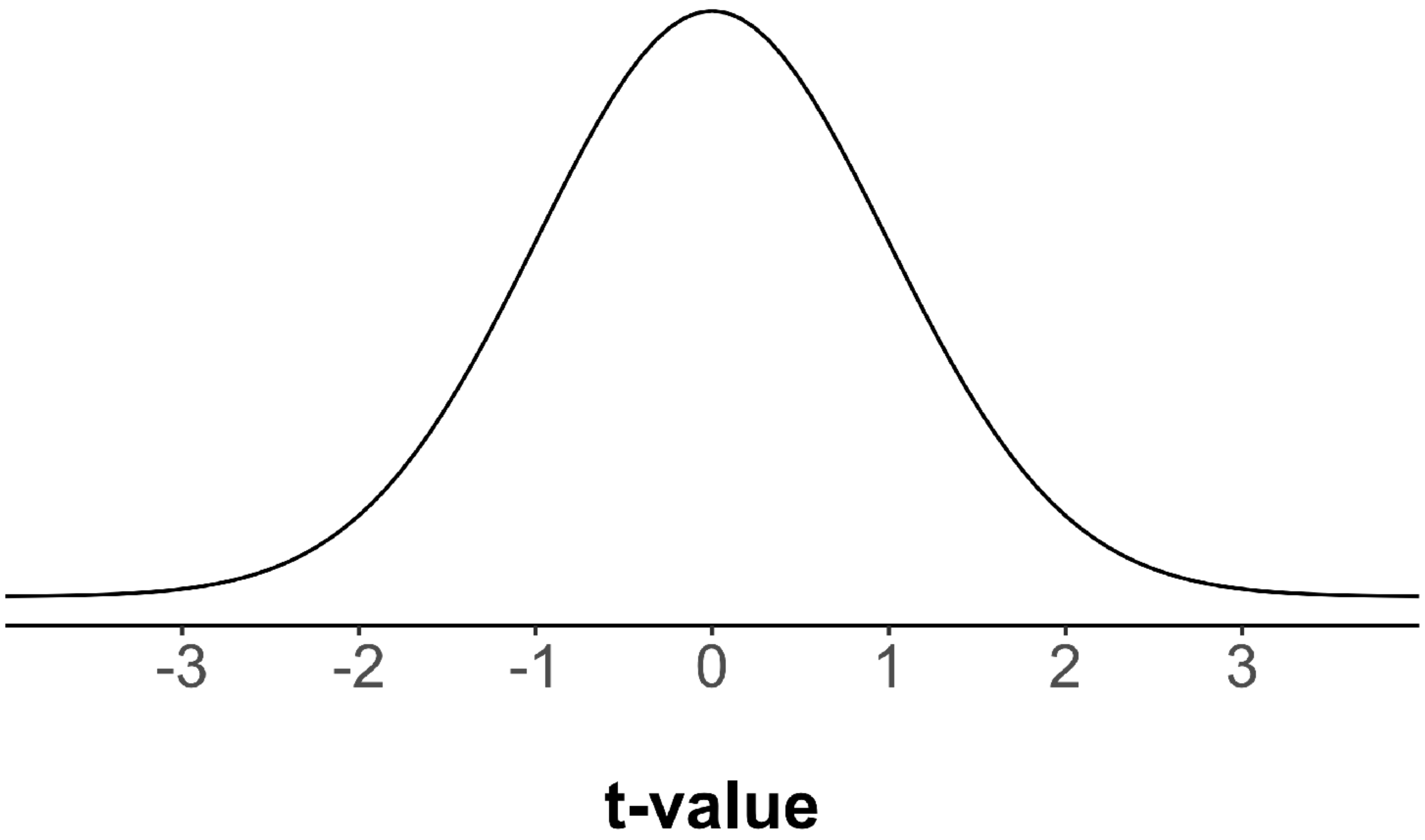
t value

$$2.7 = \frac{4 - 3.5}{\sqrt{\frac{0.2}{6}}}$$

standard error (SE)

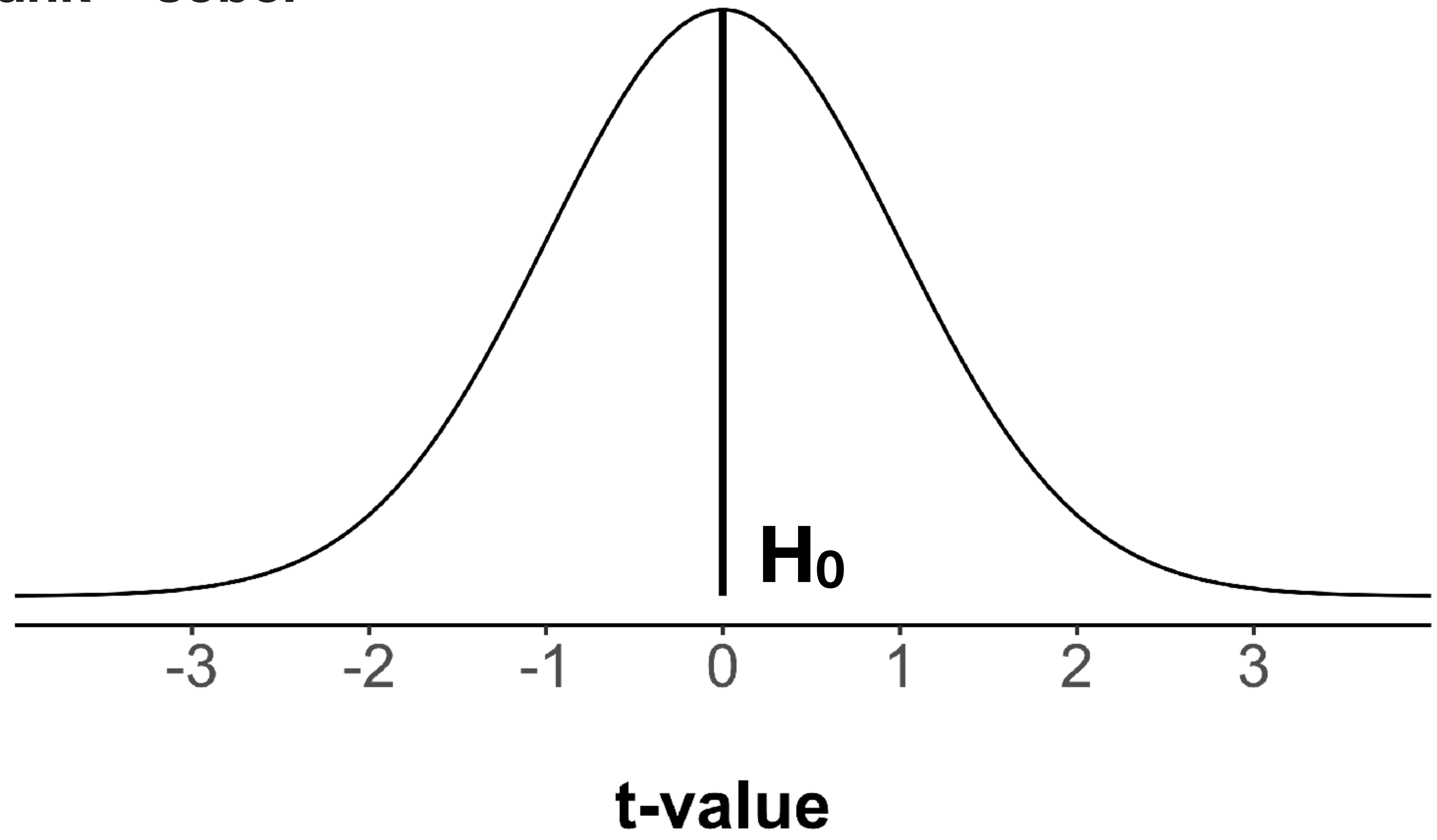
“noise”

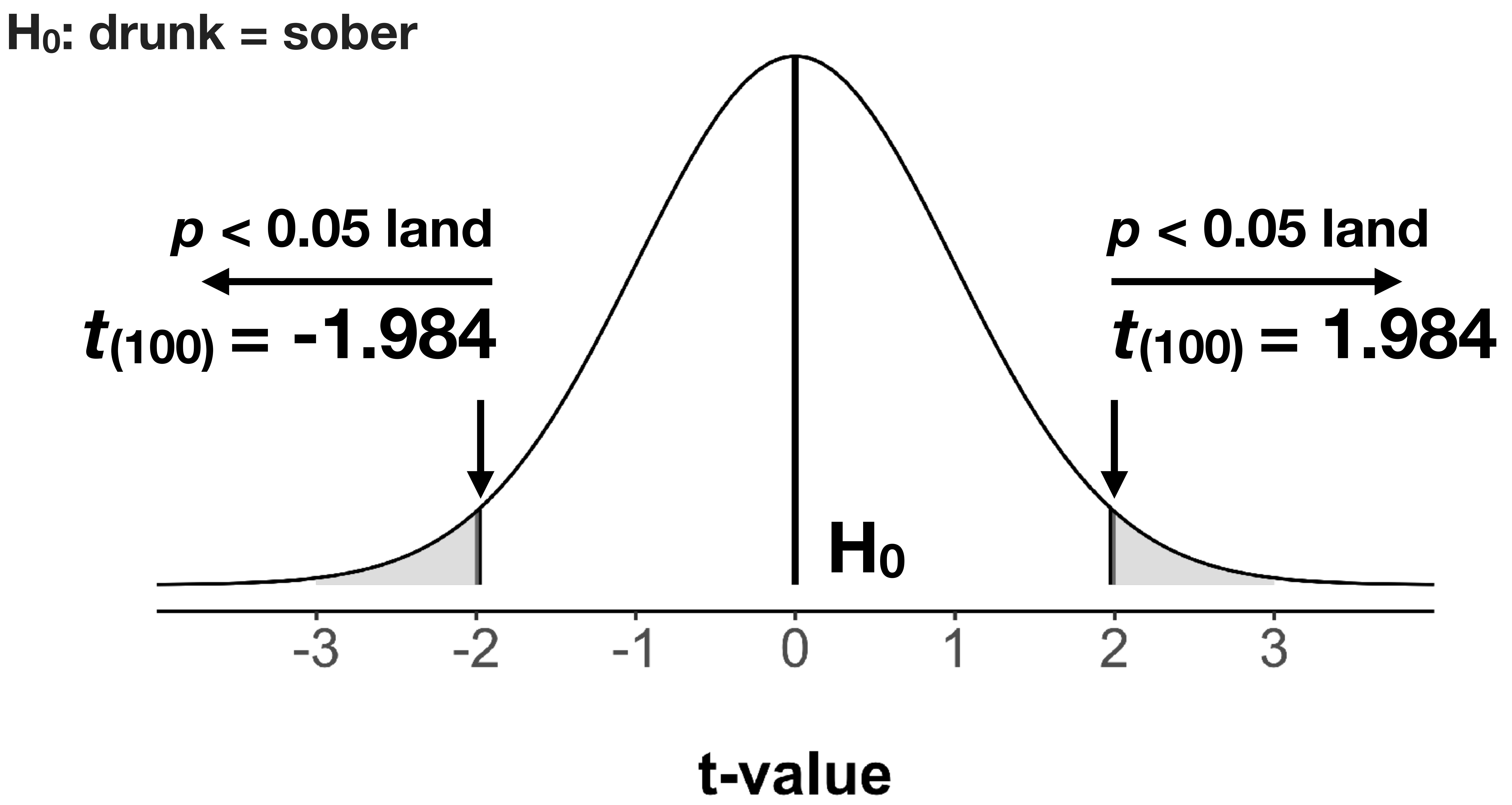
sample size



-3
-2
-1
0
1
2
3
t-value

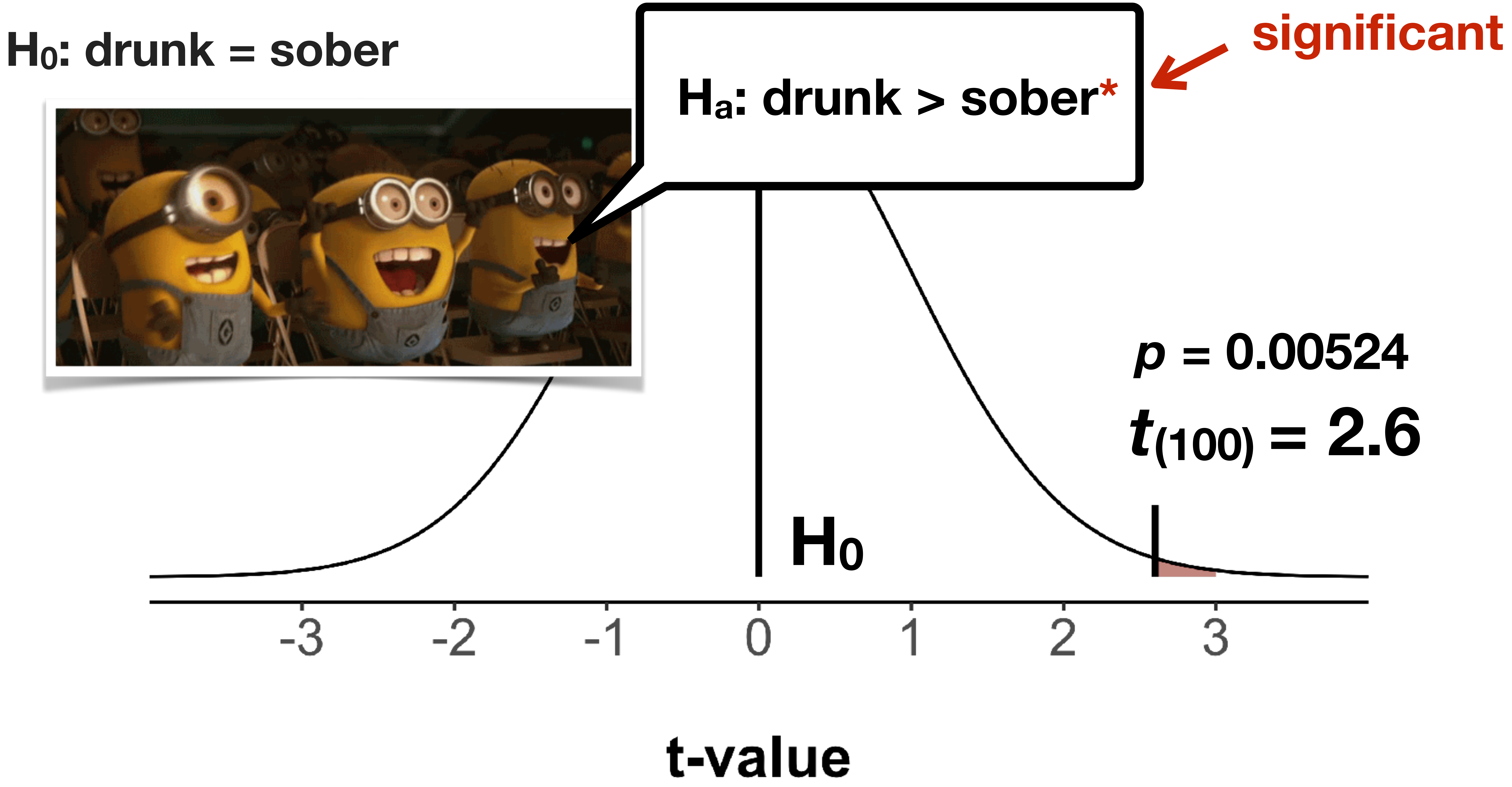
H_0: drunk = sober




H0: drunk = sober
p < 0.05 land
t(100) = -1.984
p < 0.05 land
t(100) = 1.984
H0
-3
-2
-1
0
1
2
3
t-value

drunken	sober
26	21
12	3
8	4
17	12
3	2
1	9
...	...

$$\boxed{2.6} = \frac{\boxed{6.6 - 5.6}}{\sqrt{\frac{\boxed{2.9}}{\boxed{20}}}}$$



H_0: drunk = sober
H_a: drunk > sober*
significant
$p = 0.00524$
$t_{(100)} = 2.6$
H_0
-3
-2
-1
0
1
2
3
t-value


You are DRUNK.

NHST

Null-Hypothesis Significance Testing

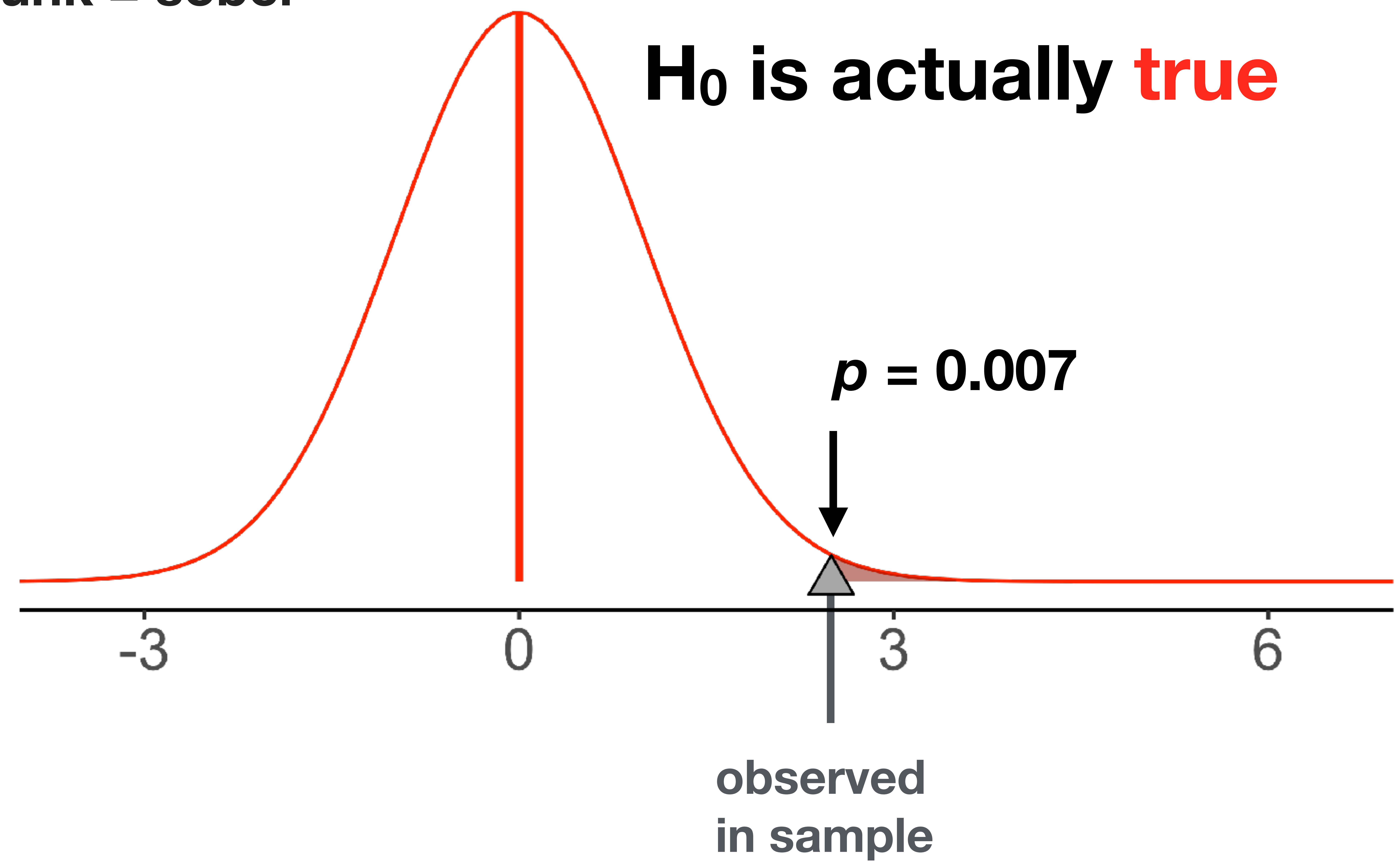
0. Set up the Alternative Hypothesis (H_1).
1. **Set up the Null-Hypothesis (H_0).**
2. **Calculate the probability of the results under H_0 (p value).**
3. **Reject H_0 when $p < 0.05$, else do not reject.**

You are **DRUNK**.

False Positive

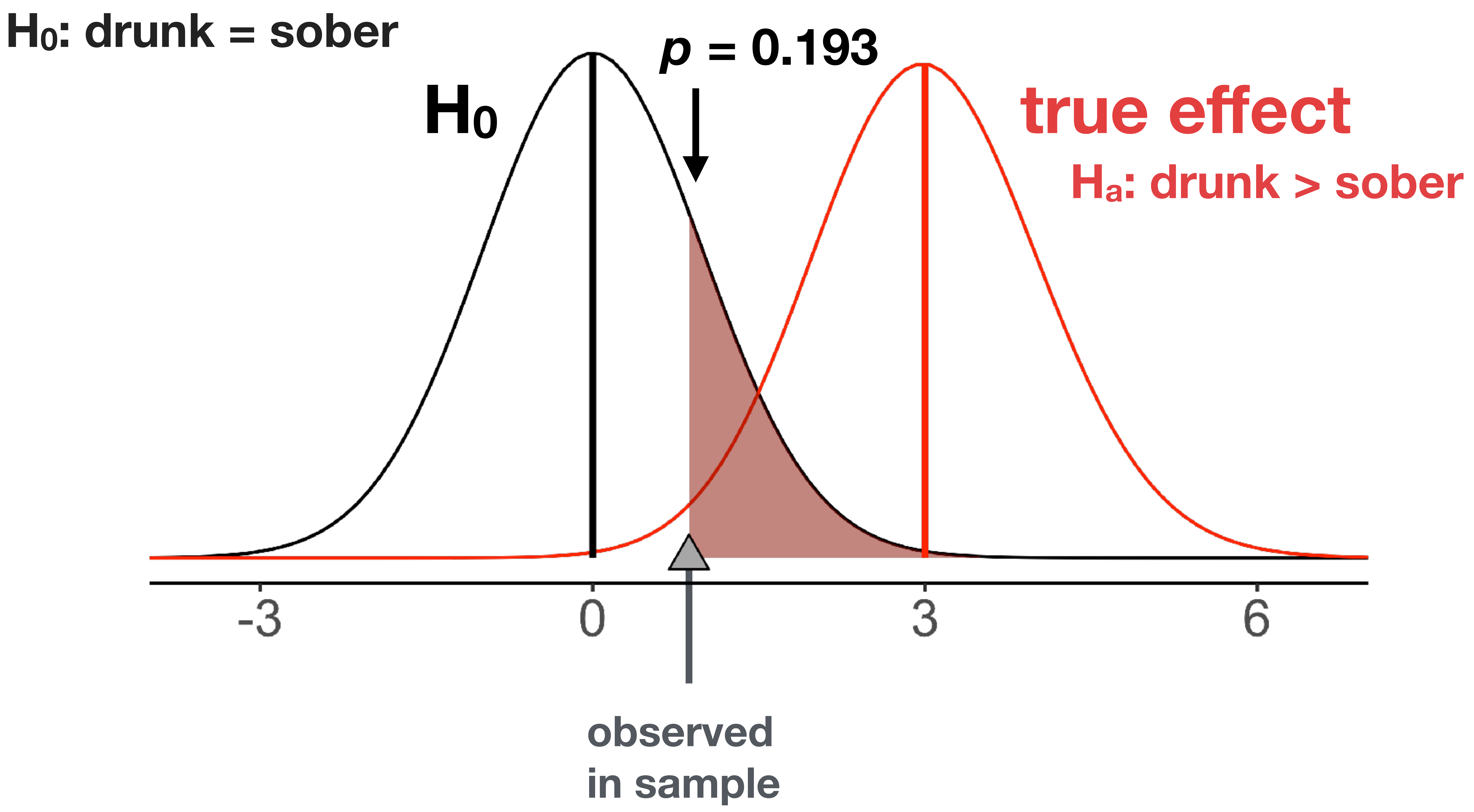
Type-I error

H$_0$: drunk = sober



False Negative

Type-II error


H_0: drunk = sober
p = 0.193
H_0
true effect
H_a: drunk > sober
-3
0
3
6
observed
in sample

Things that can go wrong

Type-I error	Erroneously **rejecting** the null
Type-II error	Erroneously **failing to** reject the null

Things that can go wrong

Type-I error Erroneously **rejecting** the null

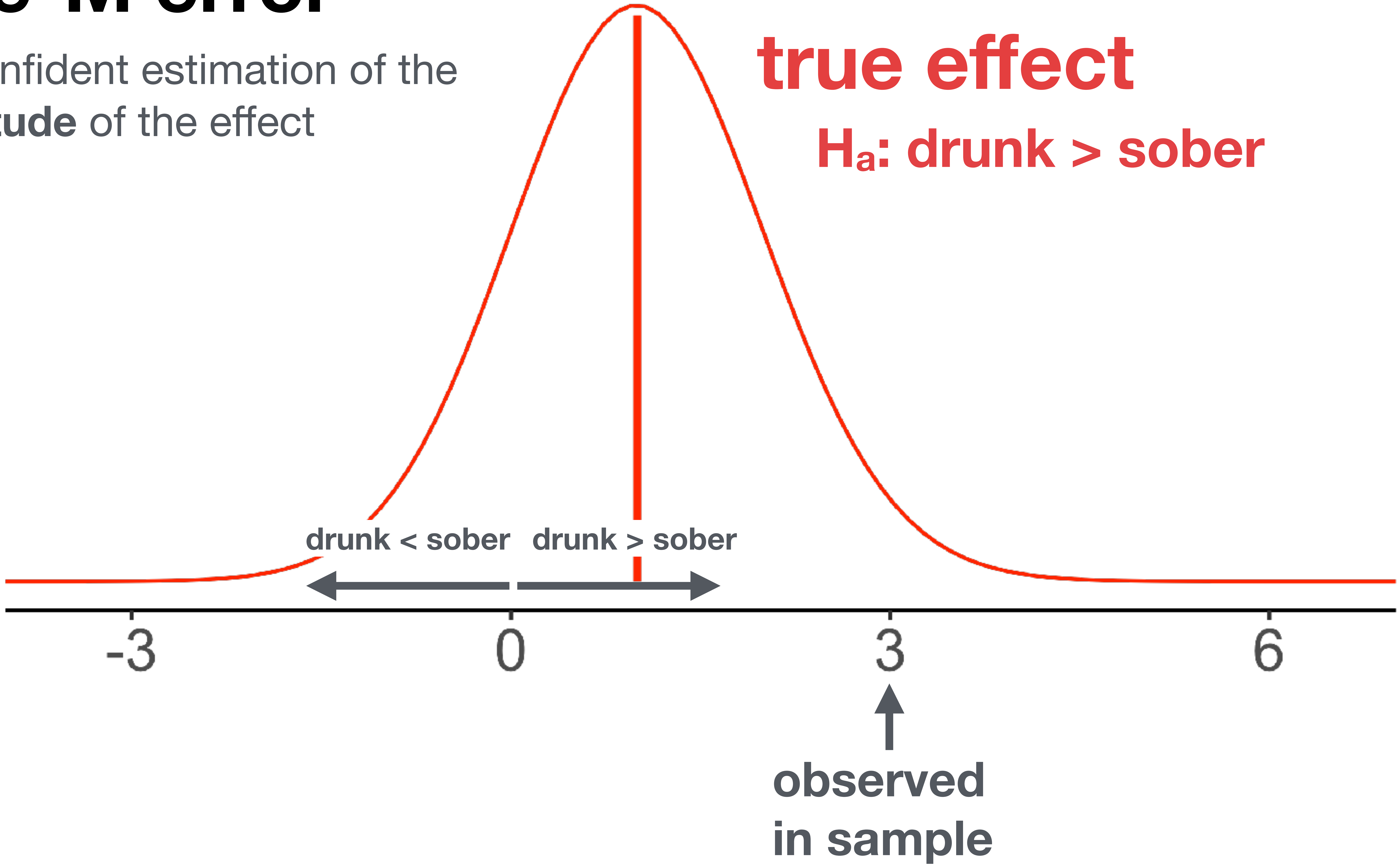
Type-II error Erroneously **failing to** reject the null

Things that can go wrong

Type-I error	Erroneously **rejecting** the null
Type-II error	Erroneously **failing to** reject the null
Type-M error	Overconfident estimation of the **magnitude** of the effect
Type-S error	Overconfident estimation of the **sign** of the effect

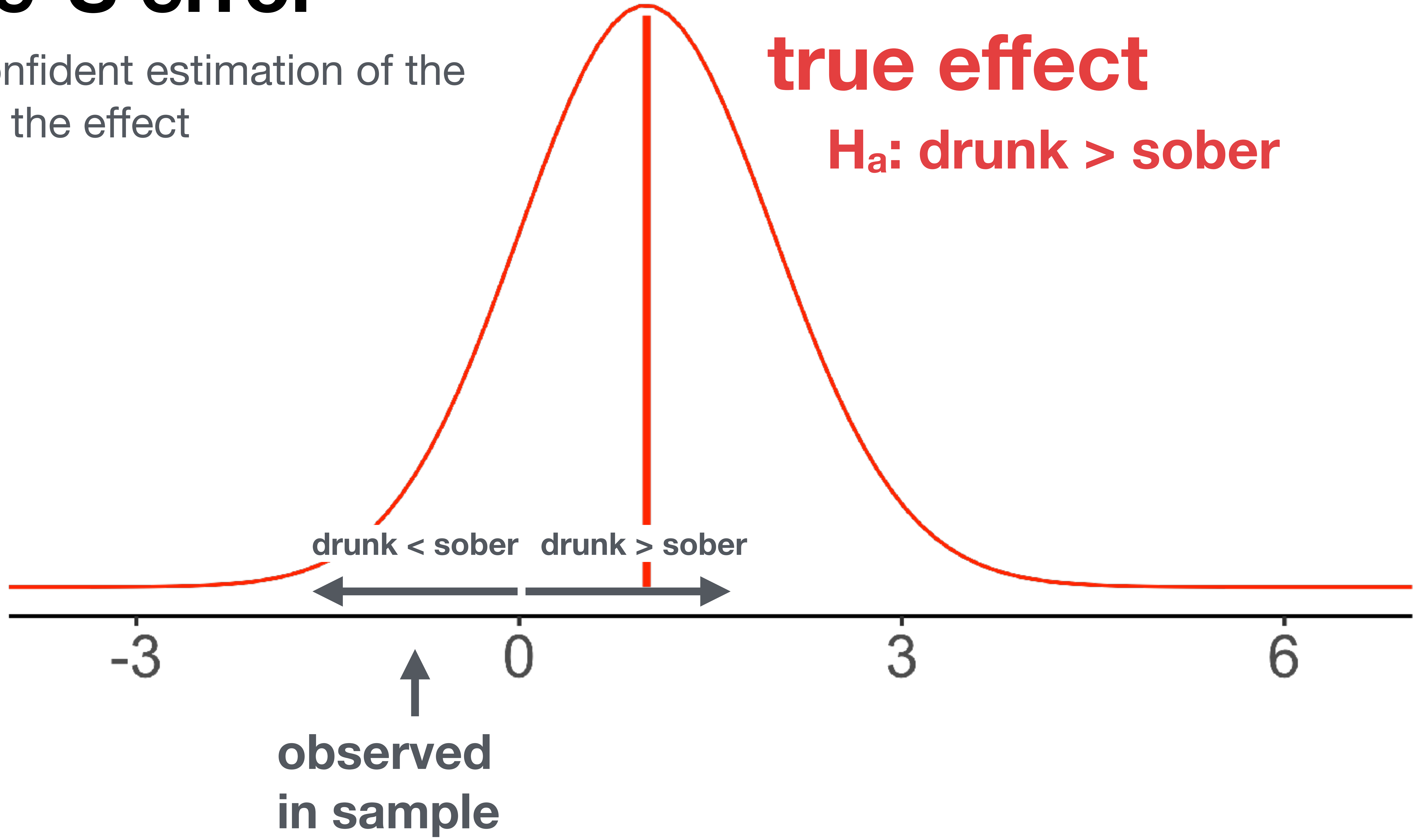
Type-M error

Overconfident estimation of the **magnitude** of the effect



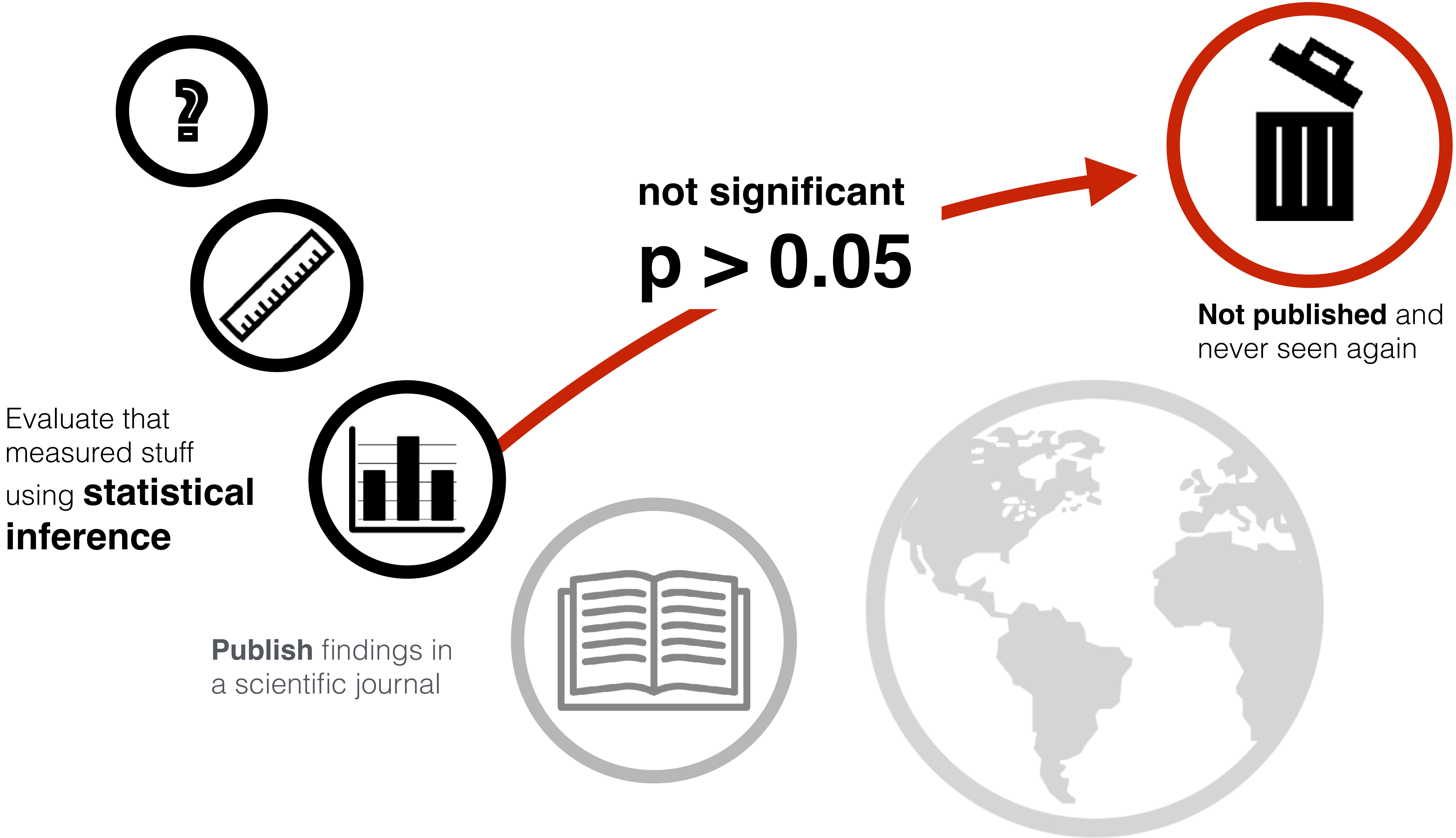
Type-S error

Overconfident estimation of the **sign** of the effect




$p < 0.05$


p < 0.05



not significant
p > 0.05
Not published and never seen again
Evaluate that measured stuff using statistical inference
Publish findings in a scientific journal

THE
p-value

p-value

The probability of the **observed result**, plus more extreme results, **if the null hypothesis was true**

- ❌ The probability of the null hypothesis
- ❌ The probability of the alternative hypothesis
- ◯
- ◯
- ◯

p-value

The probability of the **observed result**, plus more extreme results, **if the null hypothesis was true**

- ❌ The probability of the null hypothesis
- ❌ The probability of the alternative hypothesis
- ❌ If > 0.05, there is no difference between groups

p-value

The probability of the **observed result**, plus more extreme results, **if the null hypothesis was true**

- The probability of the null hypothesis
- The probability of the alternative hypothesis
- If > 0.05, there is no difference between groups
- If < 0.05, the effect is important

p-value

The probability of the **observed result**, plus more extreme results, **if the null hypothesis was true**

- The probability of the null hypothesis
- The probability of the alternative hypothesis
- If > 0.05, there is no difference between groups
- If < 0.05, the effect is important
- If < 0.05, we can conclusively answer a scientific question

p-value

The probability of the **observed result**, plus more extreme results, **if the null hypothesis was true**

- The probability of the null hypothesis
- The probability of the alternative hypothesis
- If > 0.05, there is no difference between groups
- If < 0.05, the effect is important
- If < 0.05, we can conclusively answer a scientific question

p-value

The probability of the **observed result**, plus more extreme results, **if the null hypothesis was true**

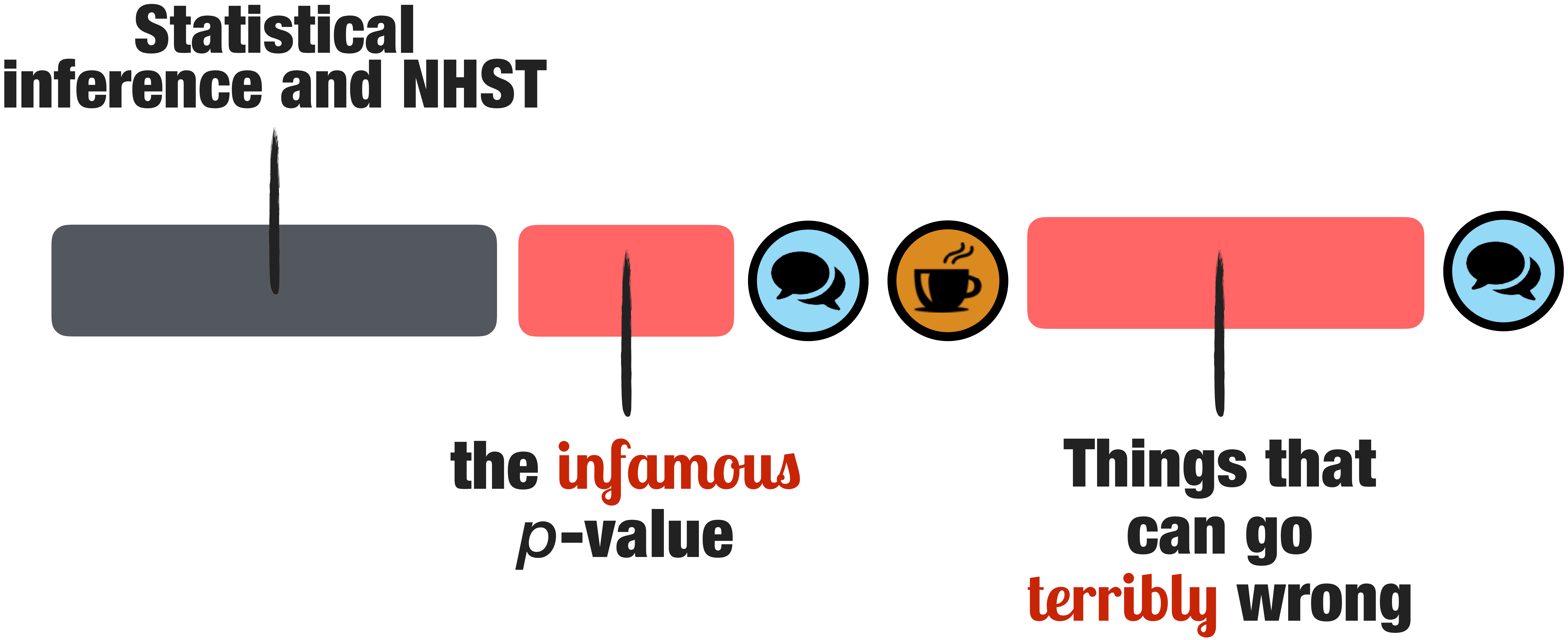
- ✗ The probability of the null hypothesis
- ✗ The probability of the alternative hypothesis
- ✗ If > 0.05, there is no difference between groups
- ✗ If < 0.05, the effect is important
- ✗ If < 0.05, we can conclusively answer a scientific question

Unlucky sampling

https://troettger.shinyapps.io/sample_away/

What are your questions?


BREAK - 10min


Statistical inference and NHST
the infamous p-value
Things that can go terribly wrong

Why do we have to be careful interpreting p-values?

We make statistical errors

Why do we have to be careful interpreting p-values?

We make statistical errors

Our studies have not enough power

Power = 1 - Type II error

Probability of correctly failing to reject the null hypothesis

Probability of erroneously failing to reject the null

t value

$$4 = \frac{4 - 2}{\sqrt{\dfrac{6}{24}}}$$

difference between groups

“noise”

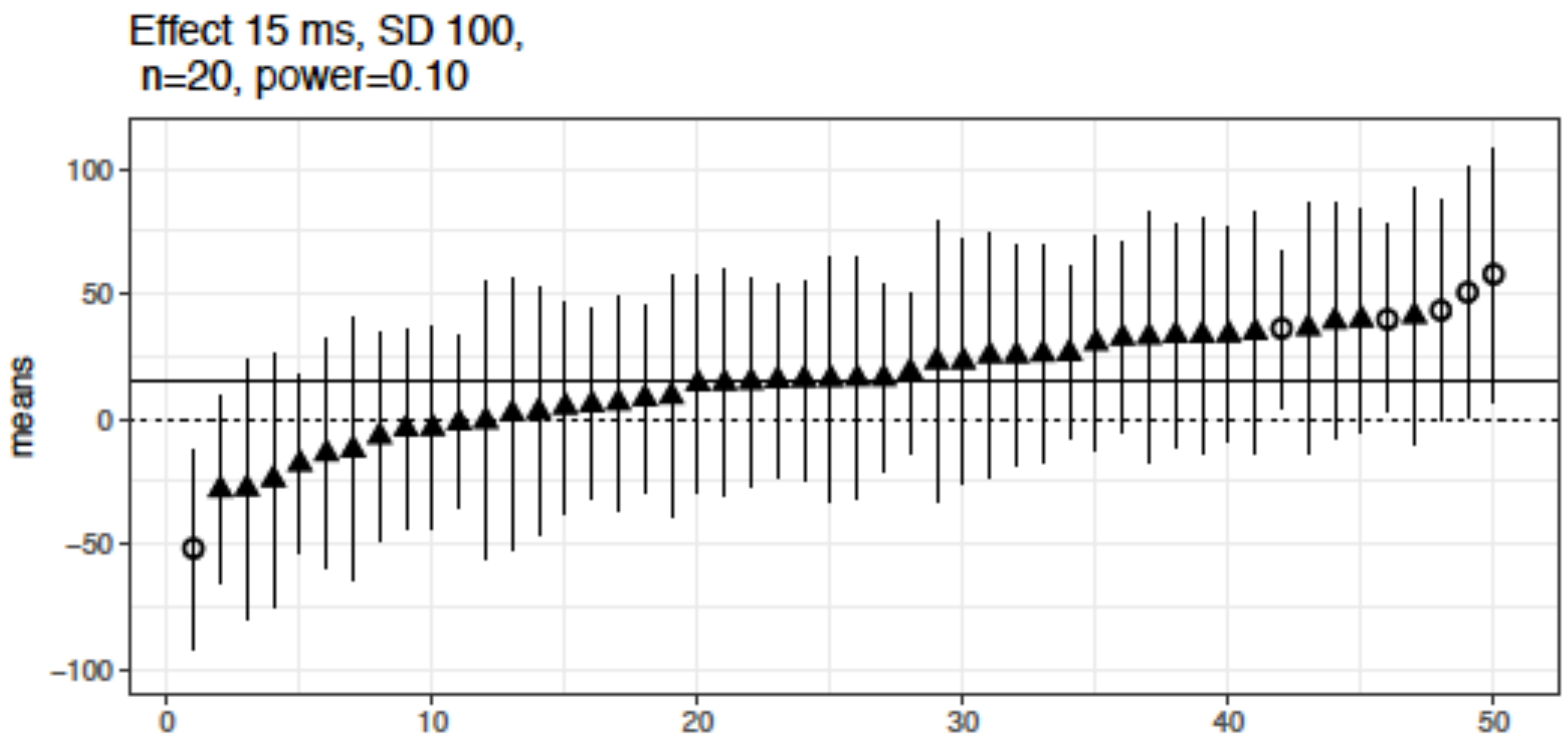
sample size

standard error (SE)

Small samples size

leads to increased

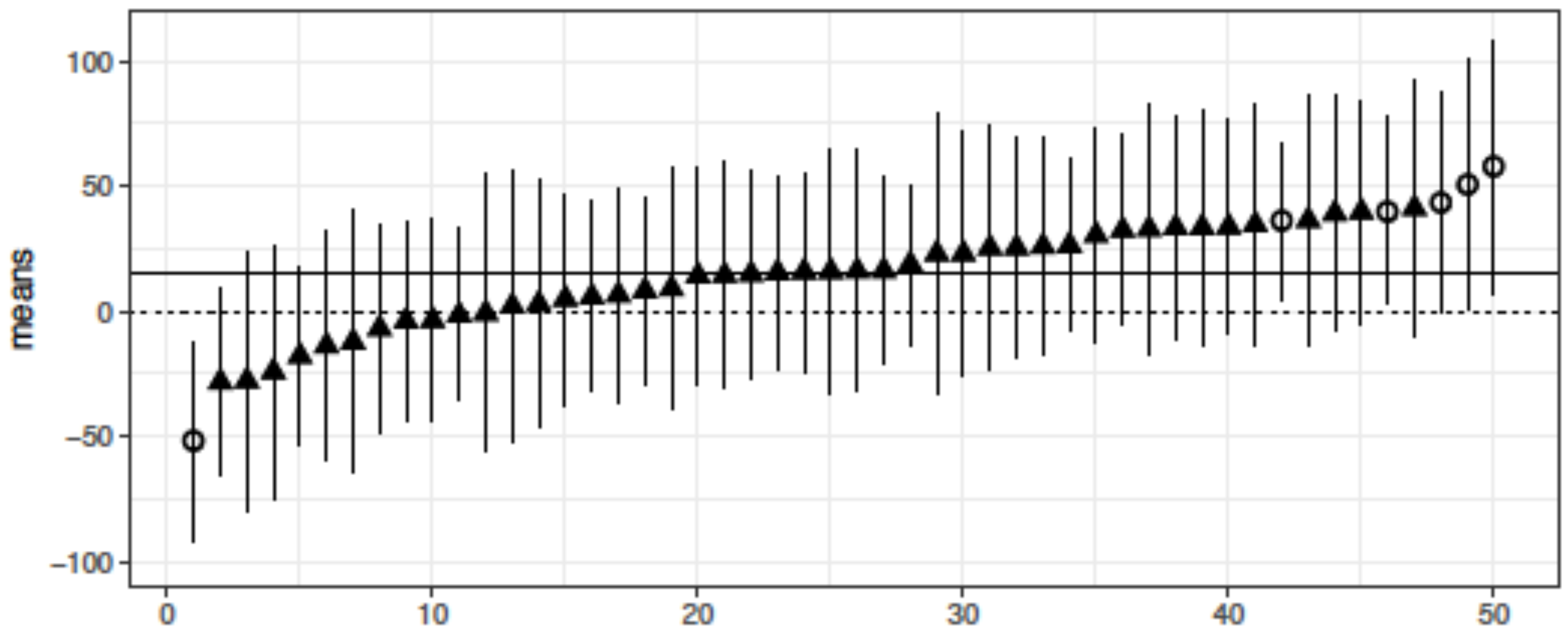
Type M & Type S errors



Vasishth et al. (2018: 152)

Small samples size leads to increased **Type M & Type S errors**

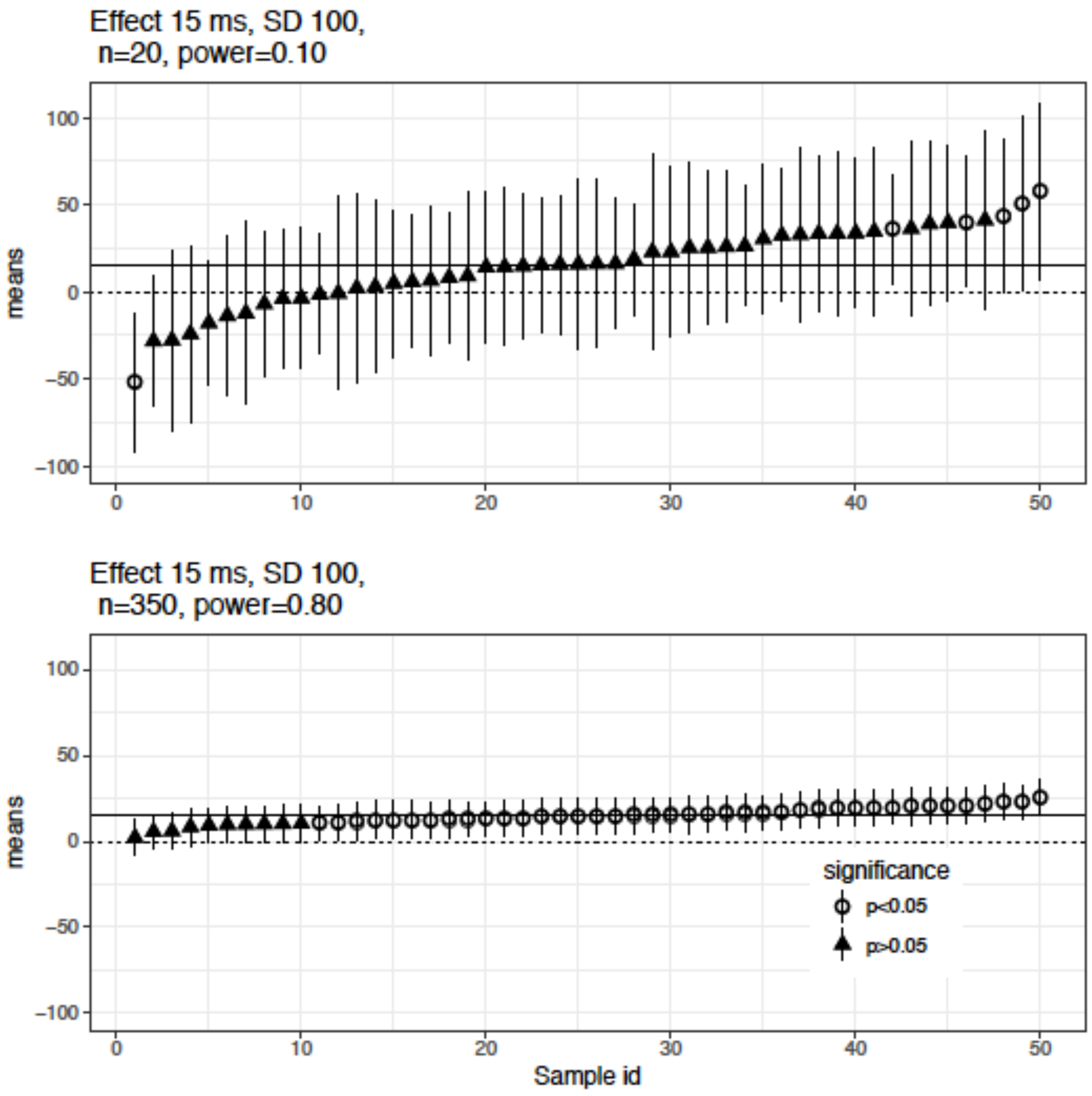
Effect 15 ms, SD 100,
n=20, power=0.10



Effect 15 ms, SD 100,
n=350, power=0.80

Vasishth et al. (2018: 152)

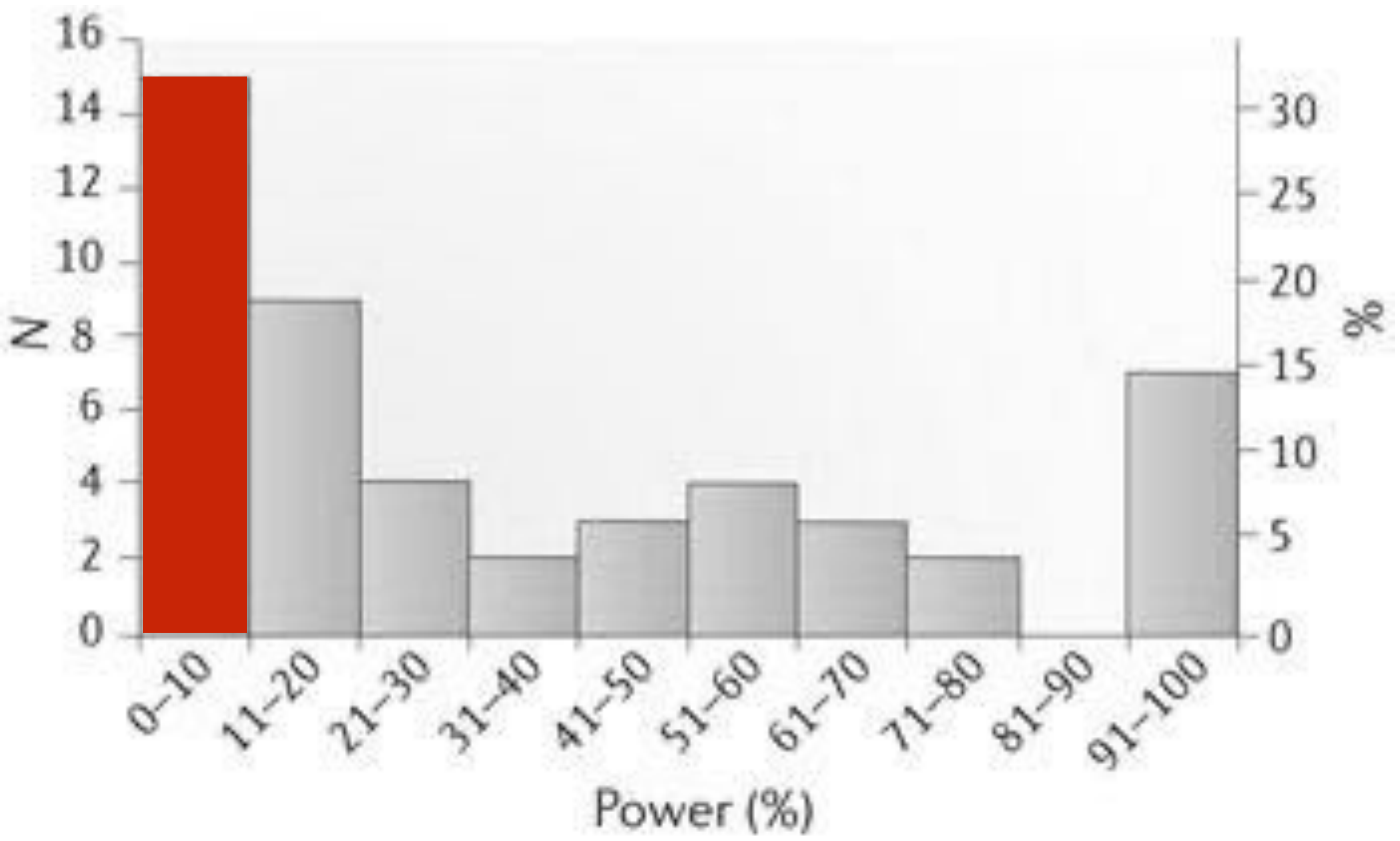
Small samples size leads to increased Type M & Type S errors



Vasishth et al. (2018: 152)

Statistical power is often low

Button et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365.



Median power = 0.21

Why do we have to be careful interpreting p-values?

We make statistical errors

Our studies have not enough power

We explore researcher degrees of freedom

The winning streak

Derren Brown

The winning streak

Derren Brown

The winning streak

Derren Brown

Multiple testing

H_1: People from Berlin are more **fashionable** than people from Osnabrück.

H_0: Berlin = Osnabrück

More V-necks
$p = 0.6$

Multiple testing

H_1: People from Berlin are more fashionable than people from Osnabrück.

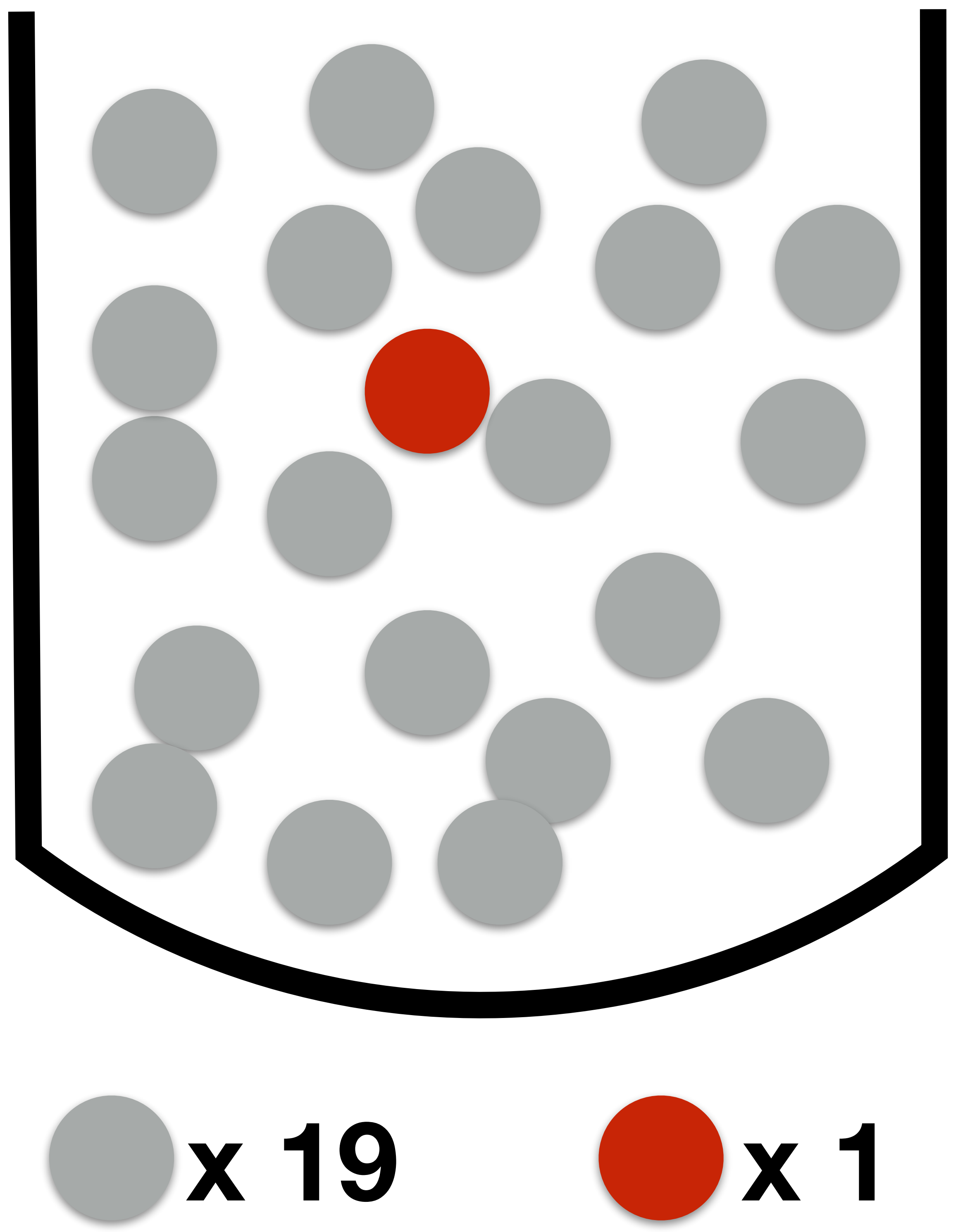
H_0: Berlin = Osnabrück

More V-necks
$p = 0.6$

More colorful
$p = 0.04$

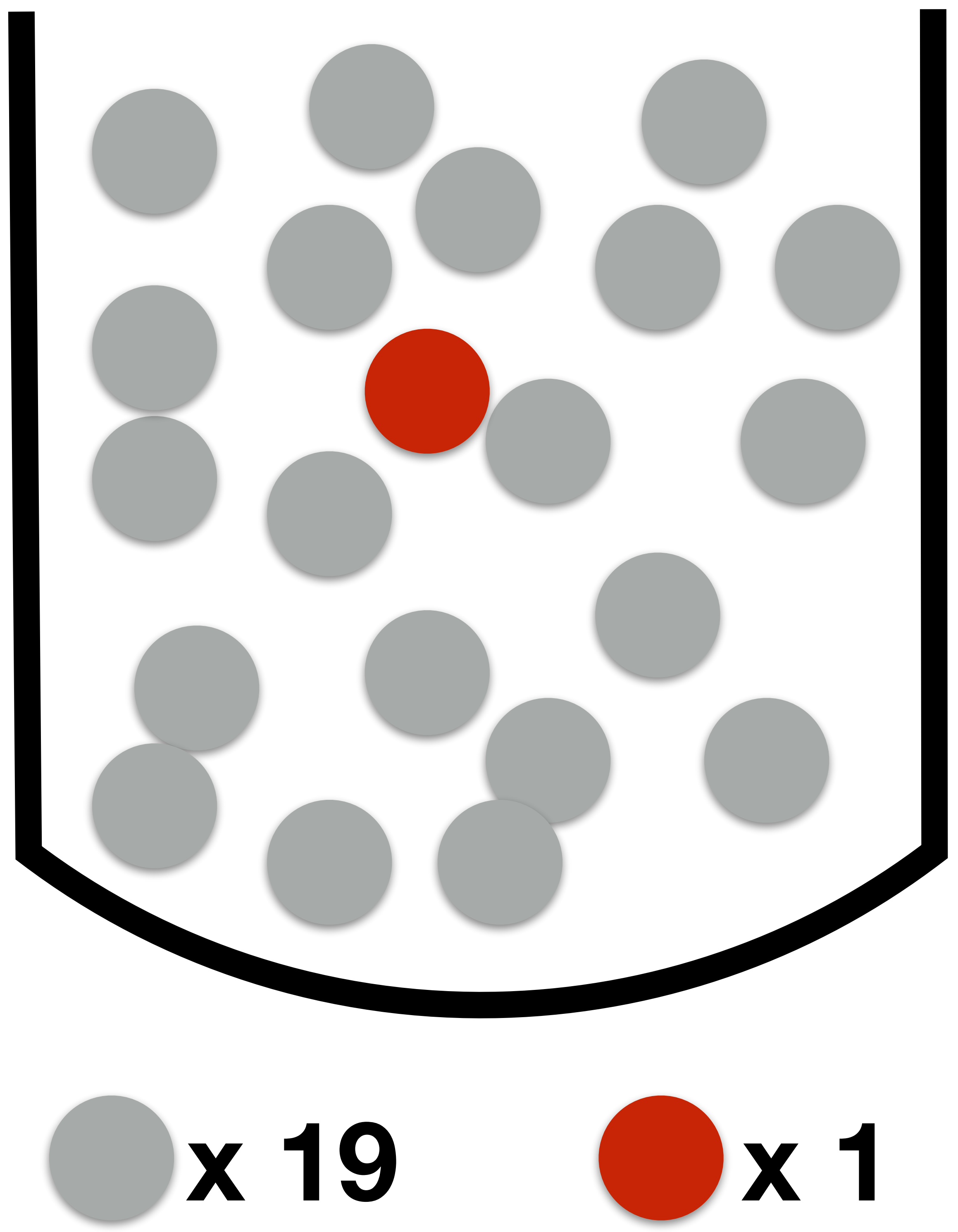
Probability of randomly pulling the red marble?

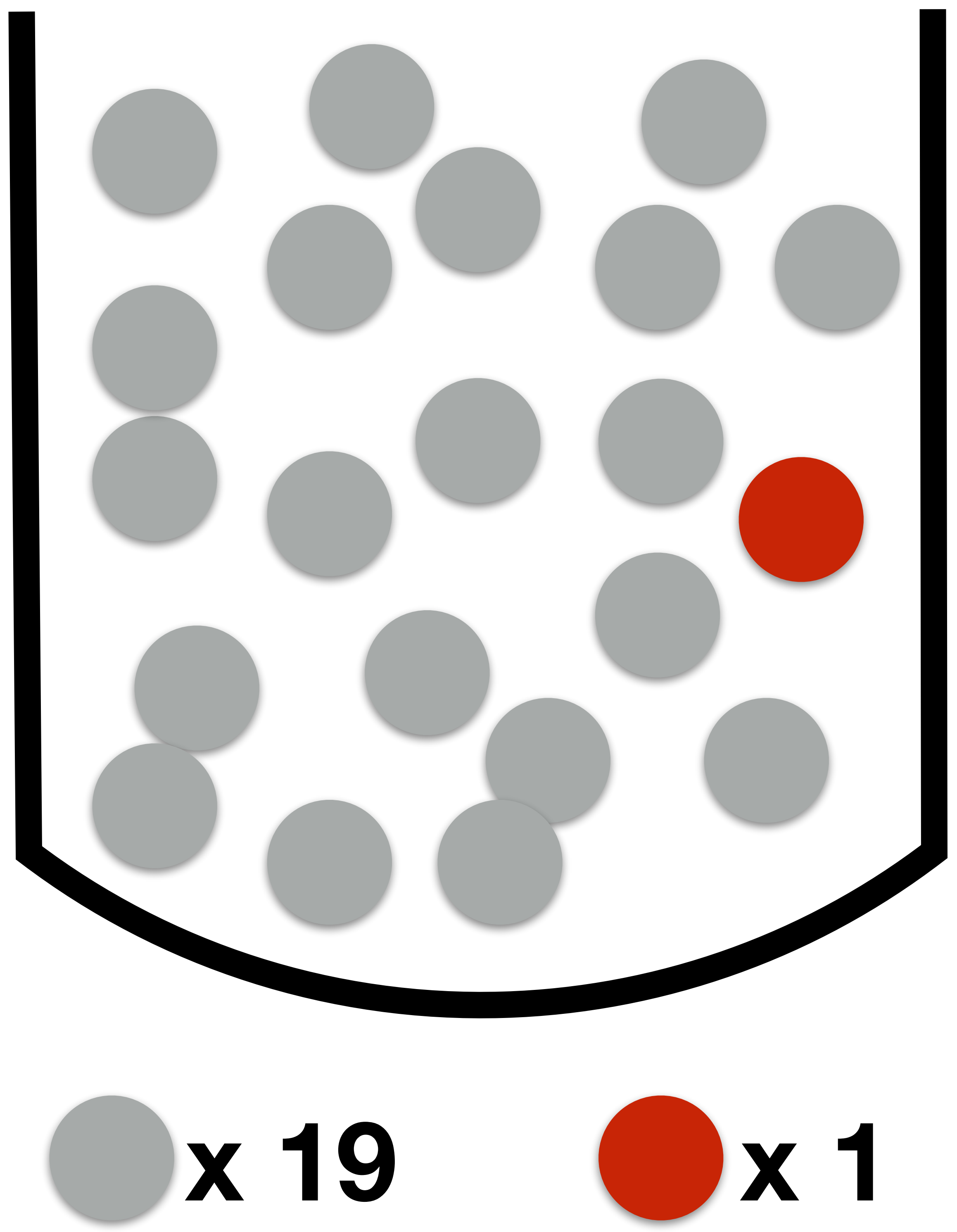
0.05



Probability of randomly pulling one red marble out of one of the bowls?

$$1 - (1 - 0.05)^2$$
$$= 0.0975$$





The interpretation of the p-value is affected by researcher degrees of freedom

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*

Andrew Gelman† and Eric Loken‡

14 Nov 2013

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

The interpretation of the p-value is affected by researcher degrees of freedom

an be a problem,
and the research

Laboratory Phonology
Journal of the Association for Laboratory Phonology

Roettger, T. B. 2019 Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1): 1, pp.1–27. DOI: https://doi.org/10.5334/labphon.147

JOURNAL ARTICLE

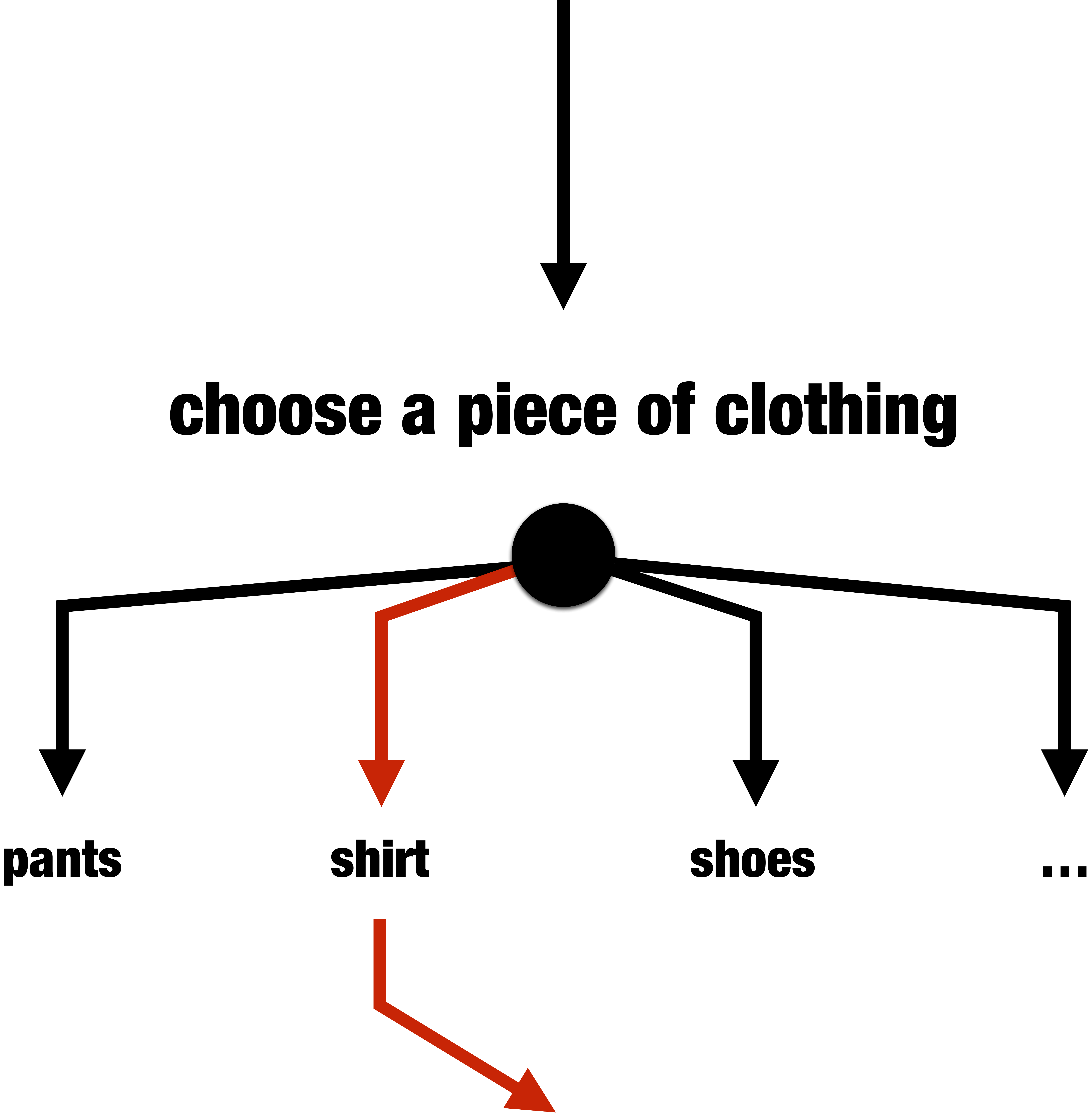
Researcher degrees of freedom in phonetic research

Timo B. Roettger
Department of Linguistics, Northwestern University, Evanston, IL, US
timo.b.roettger@gmail.com

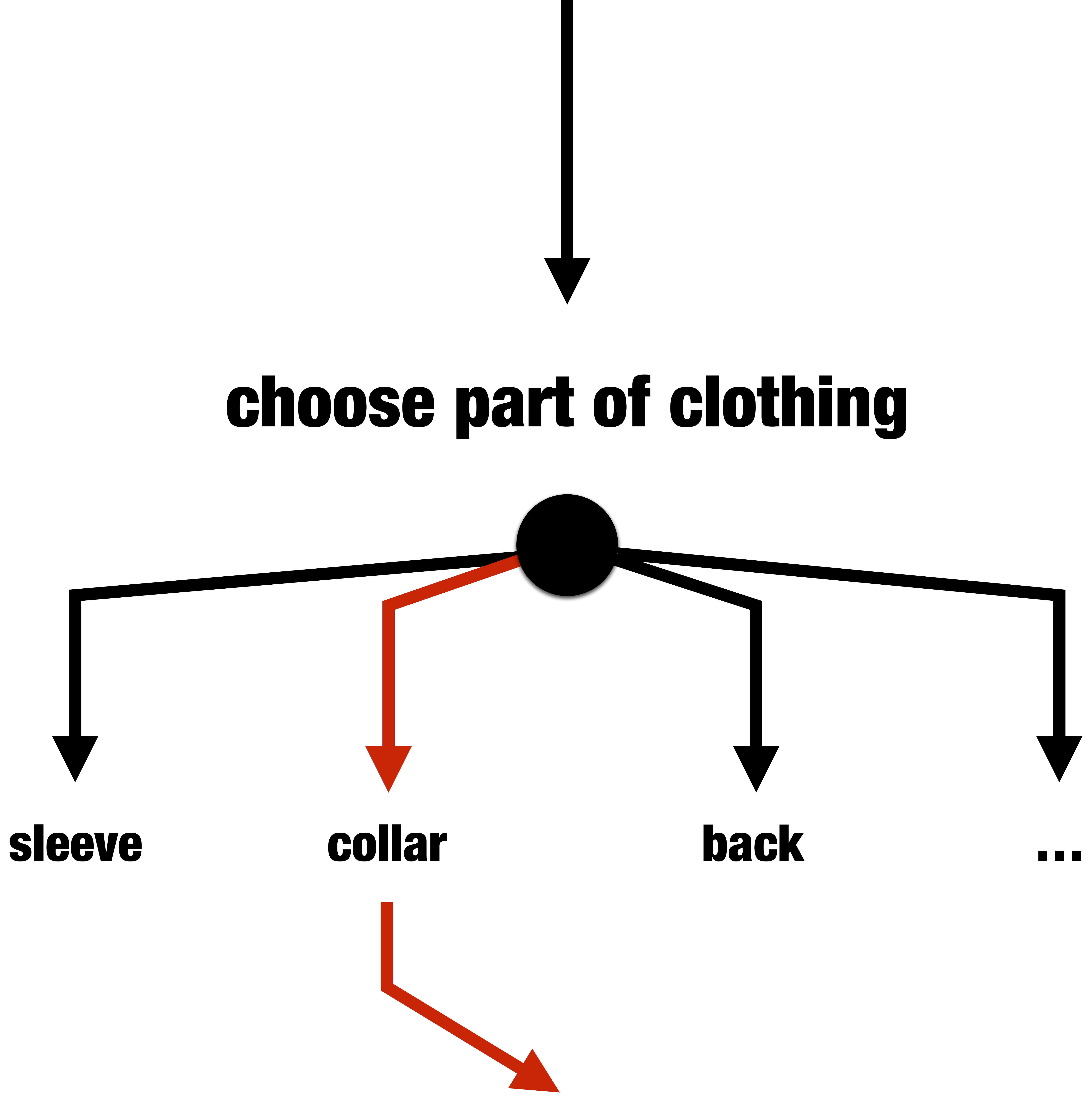
s.nav
32

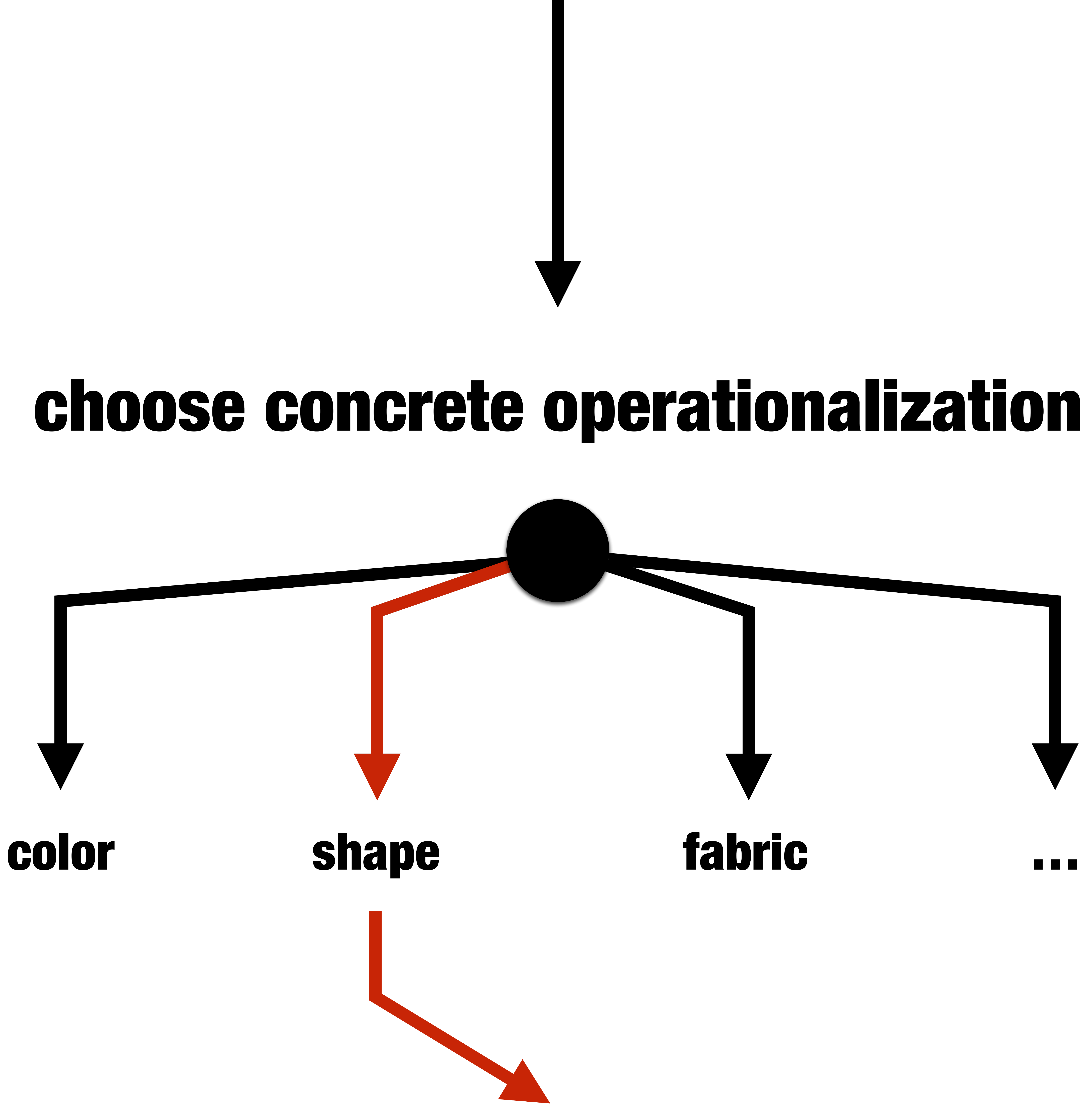
False-Po
Flexibilit
Allows P

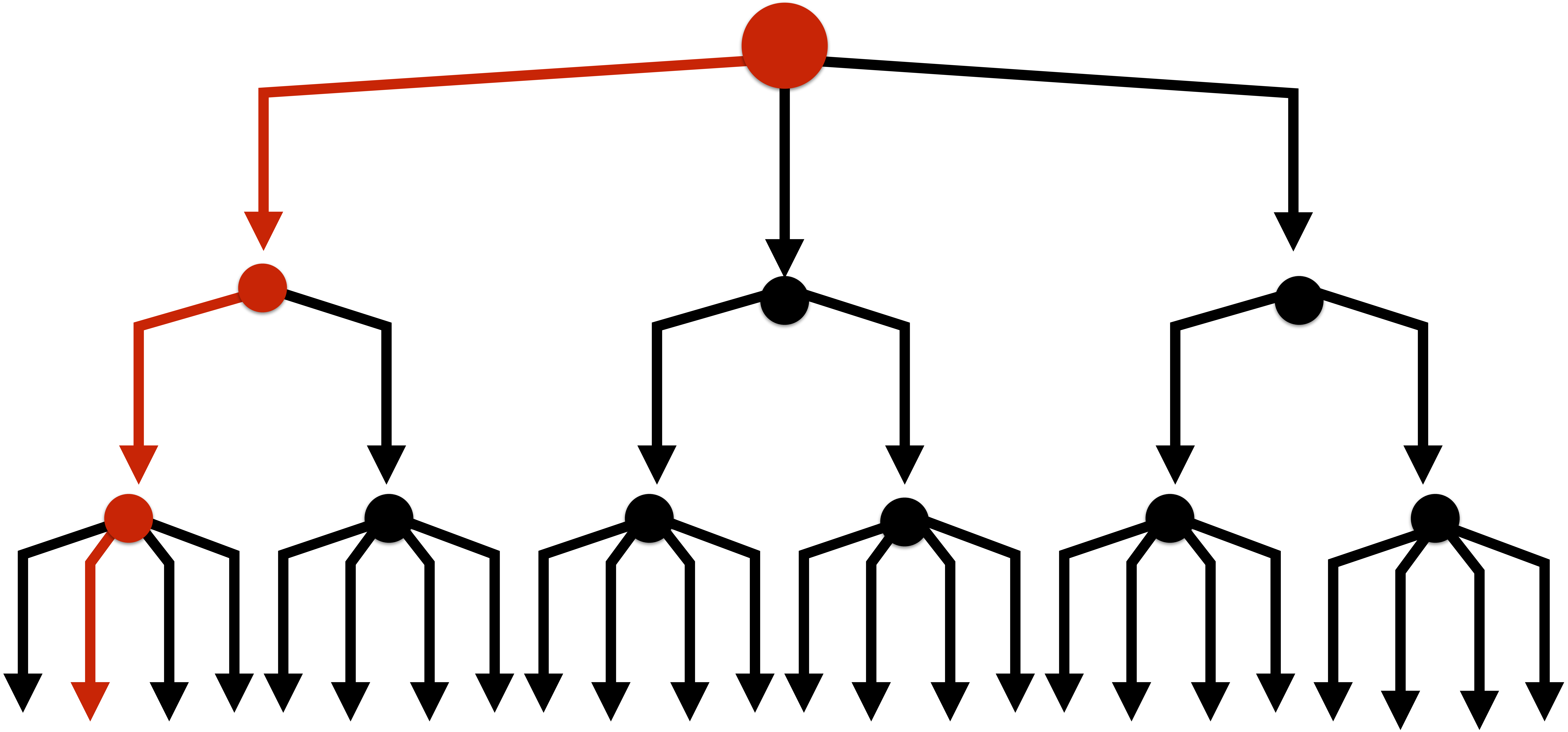
Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

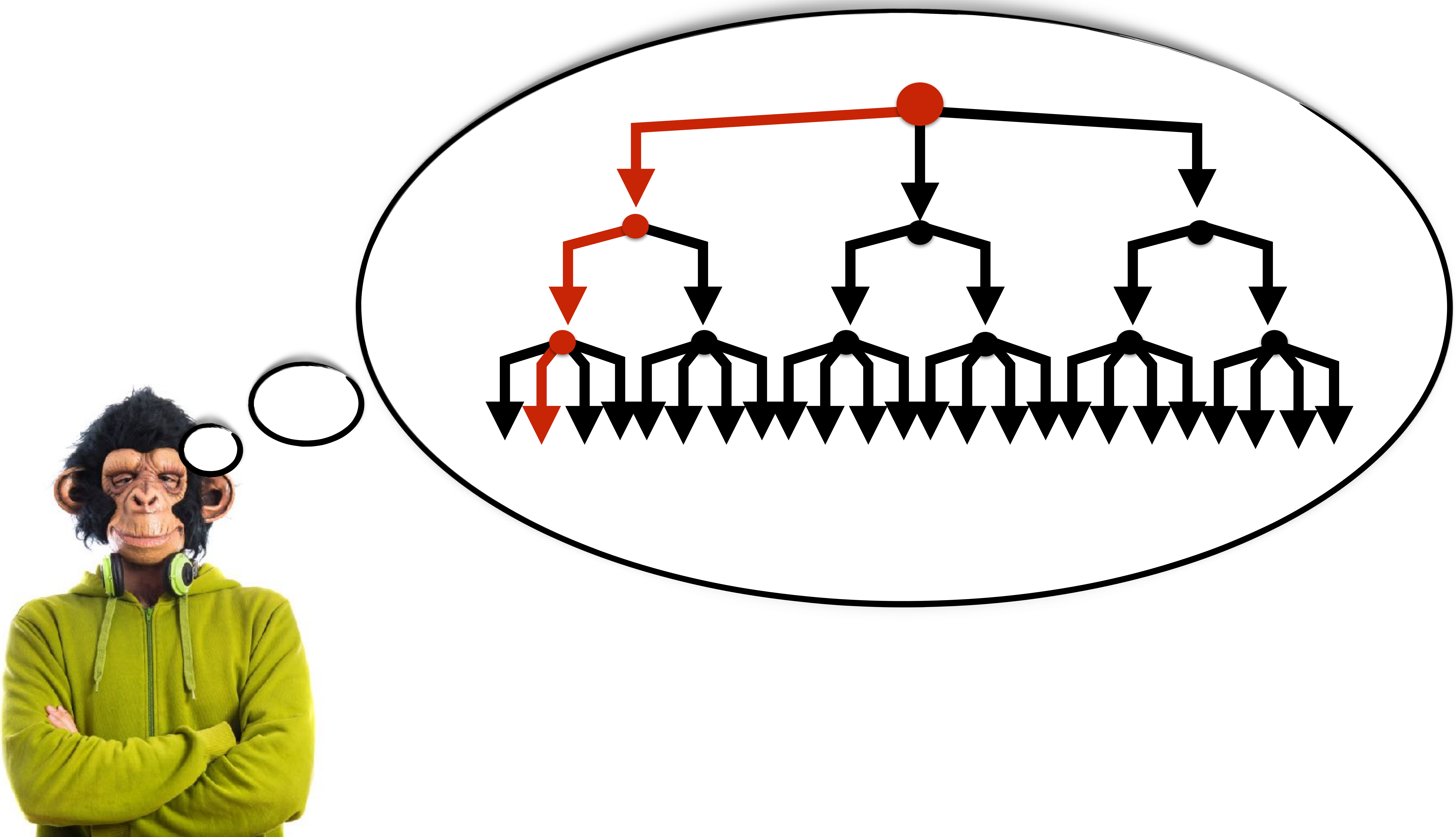


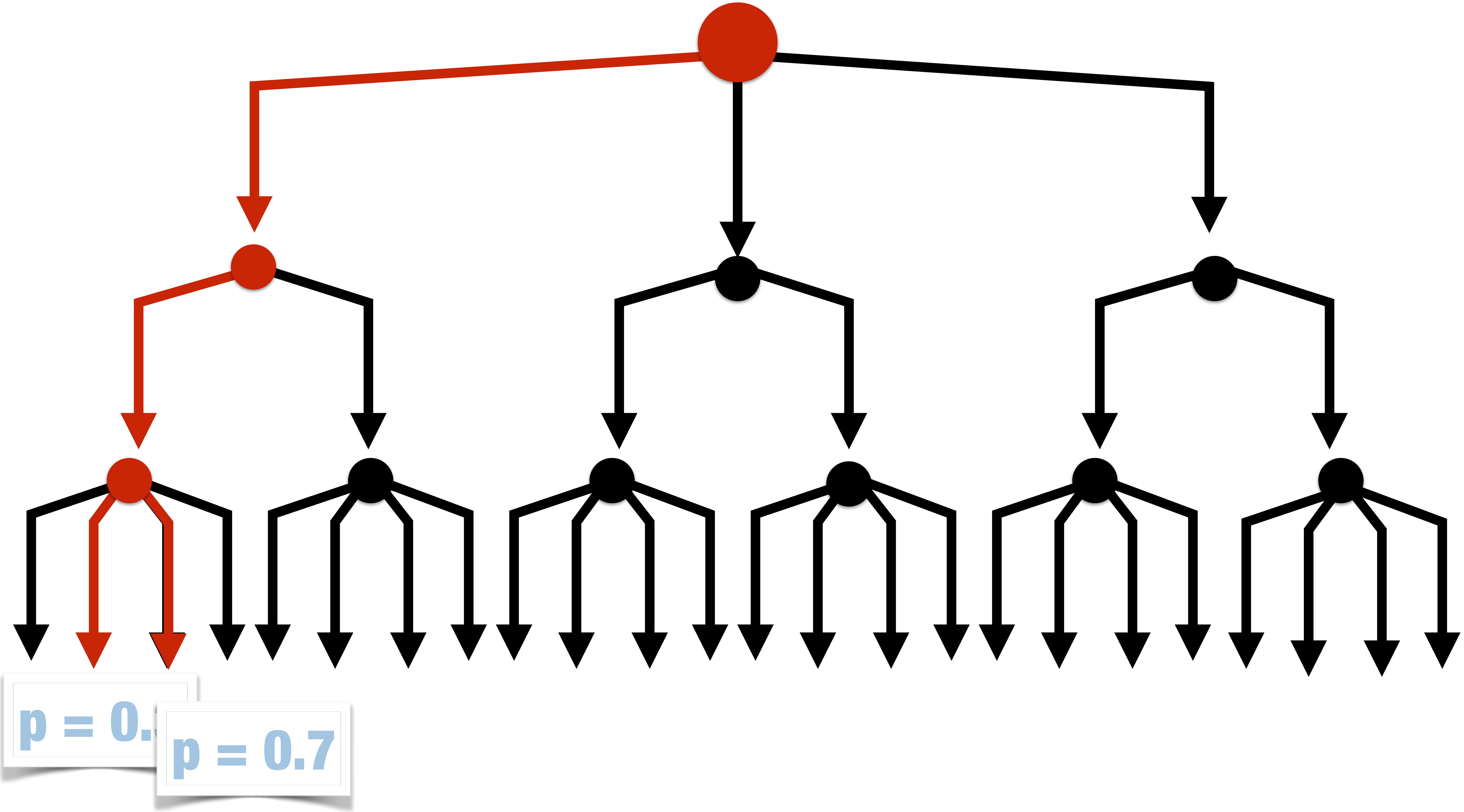
choose a piece of clothing
pants
shirt
shoes
...

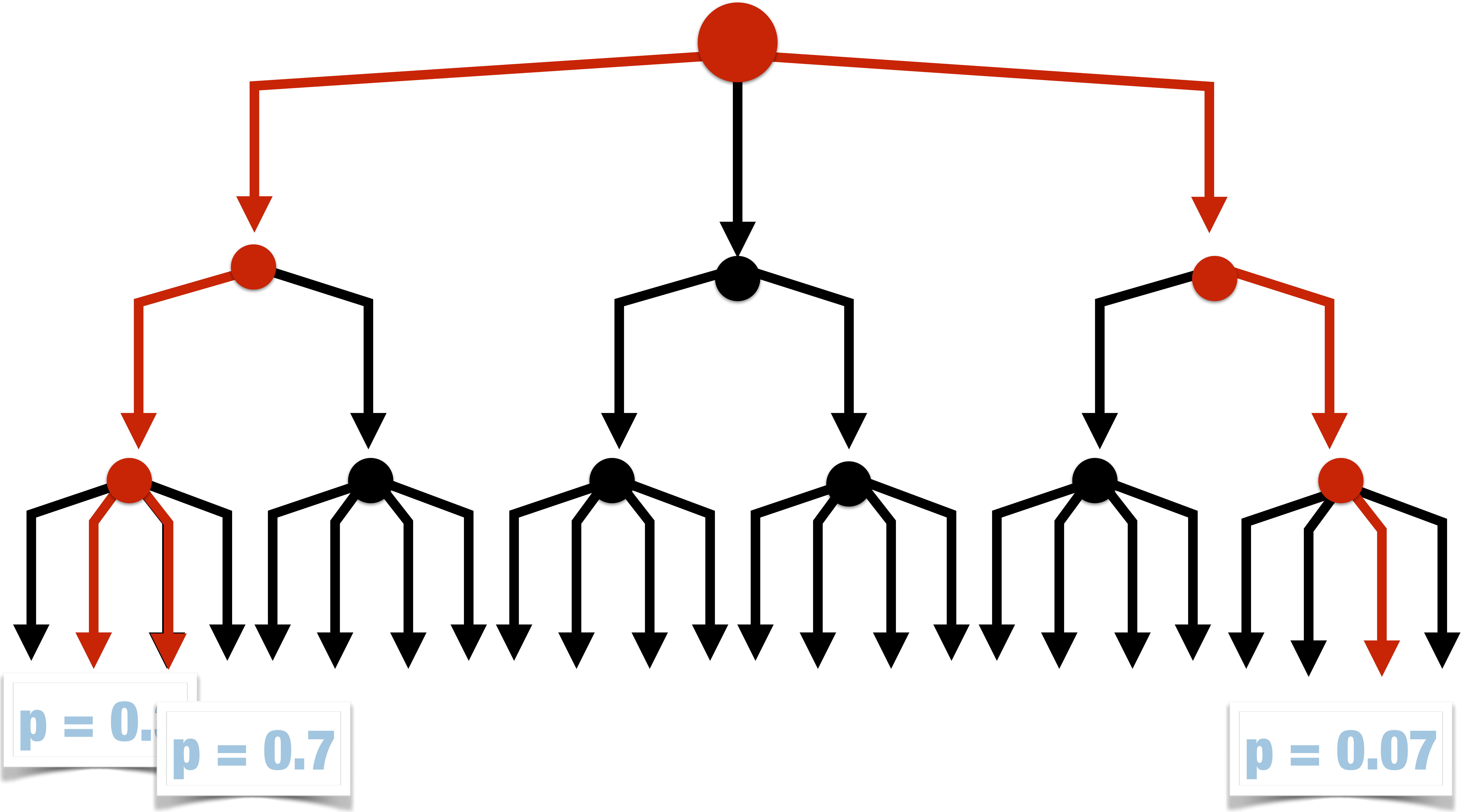

choose part of clothing
sleeve
collar
back
...

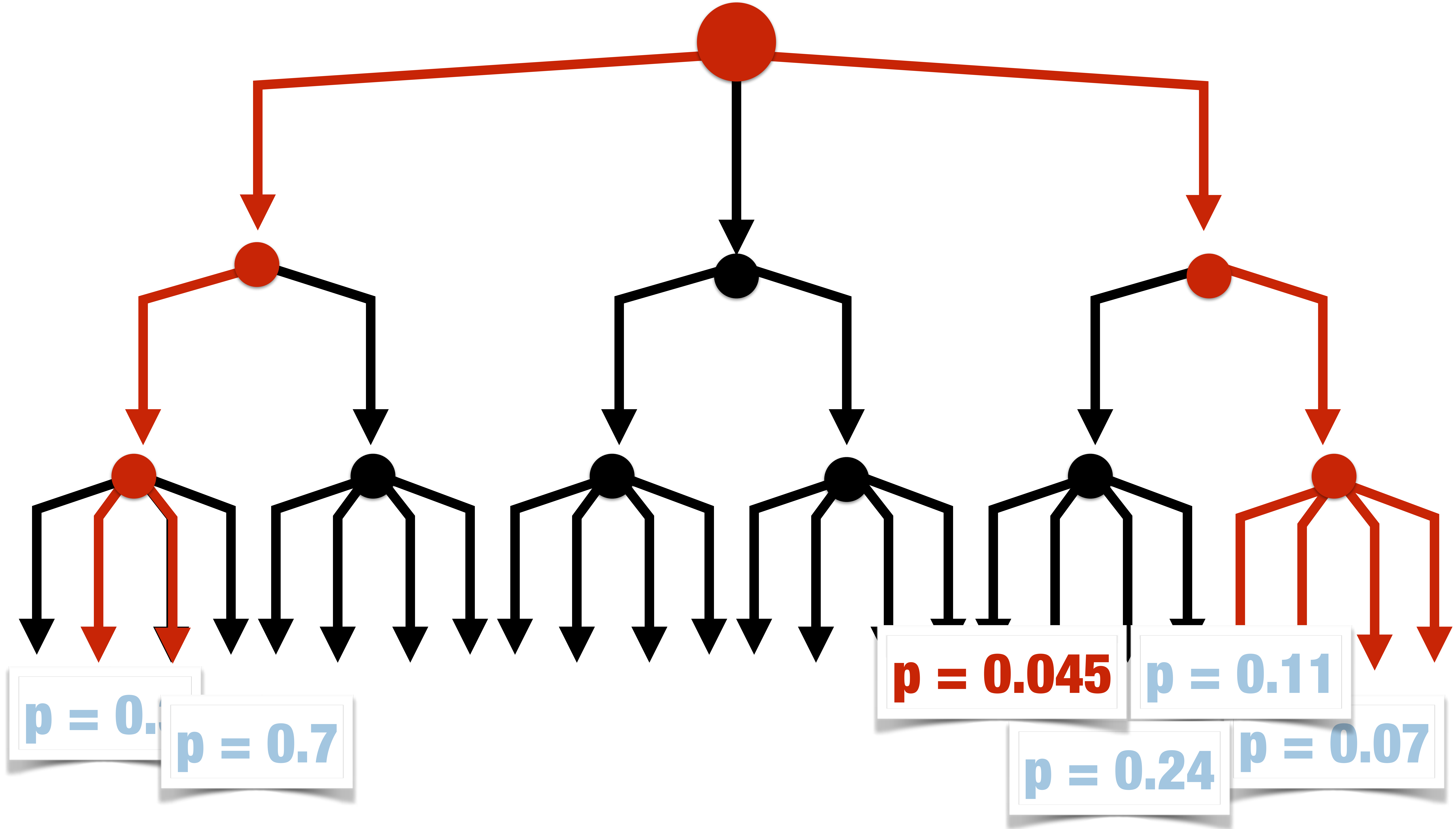

choose concrete operationalization
color
shape
fabric
...



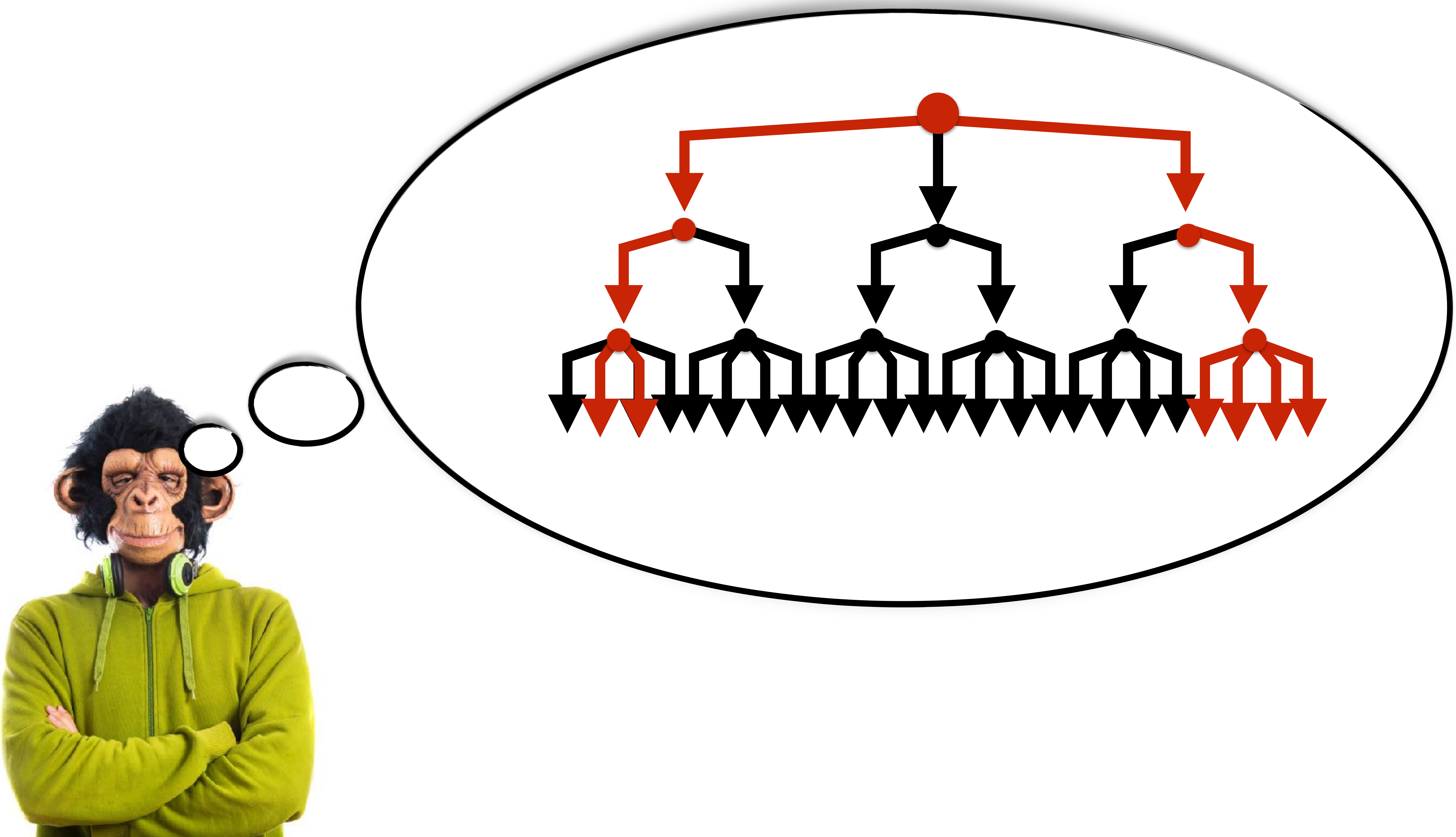


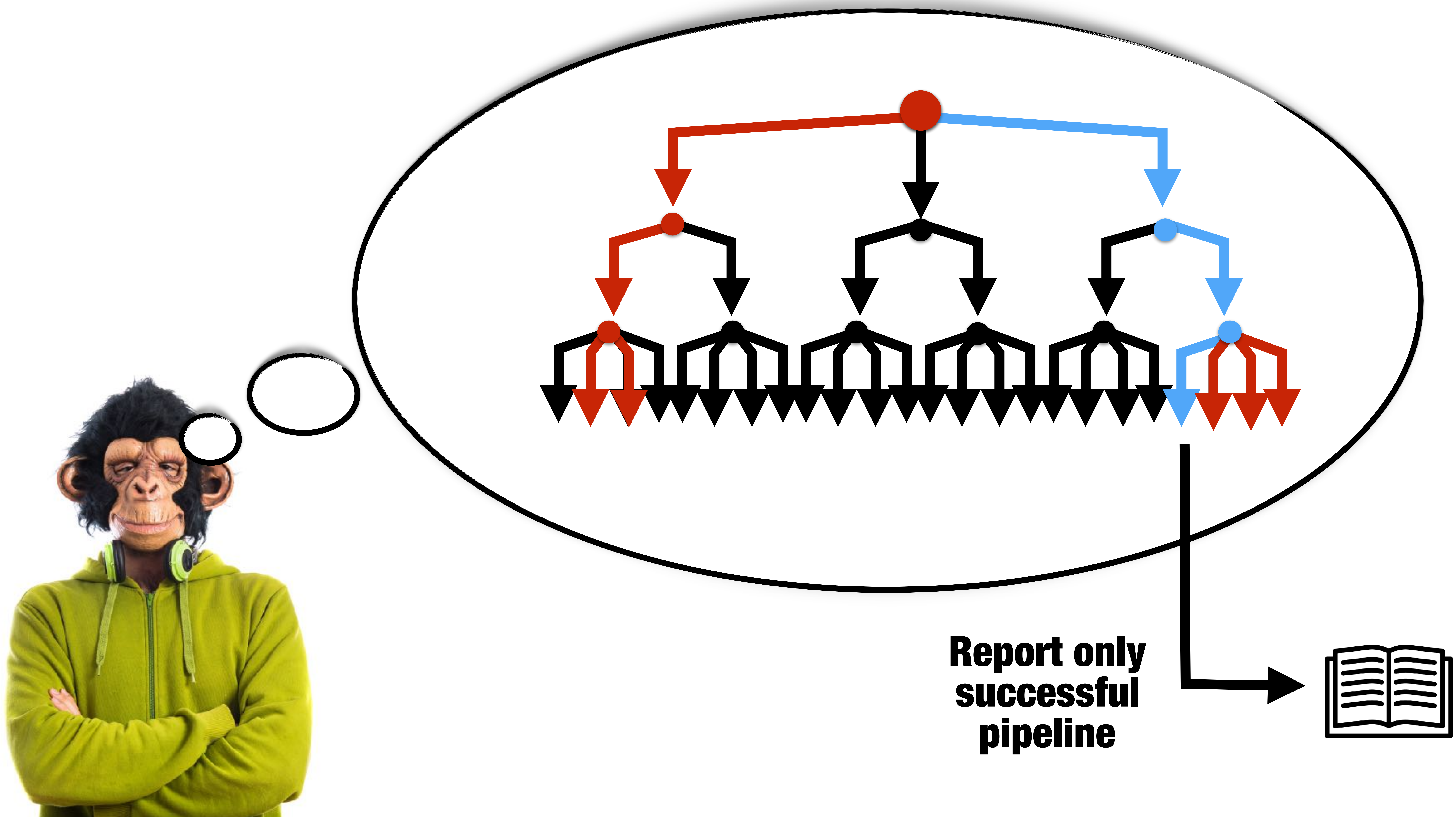

p = 0.7

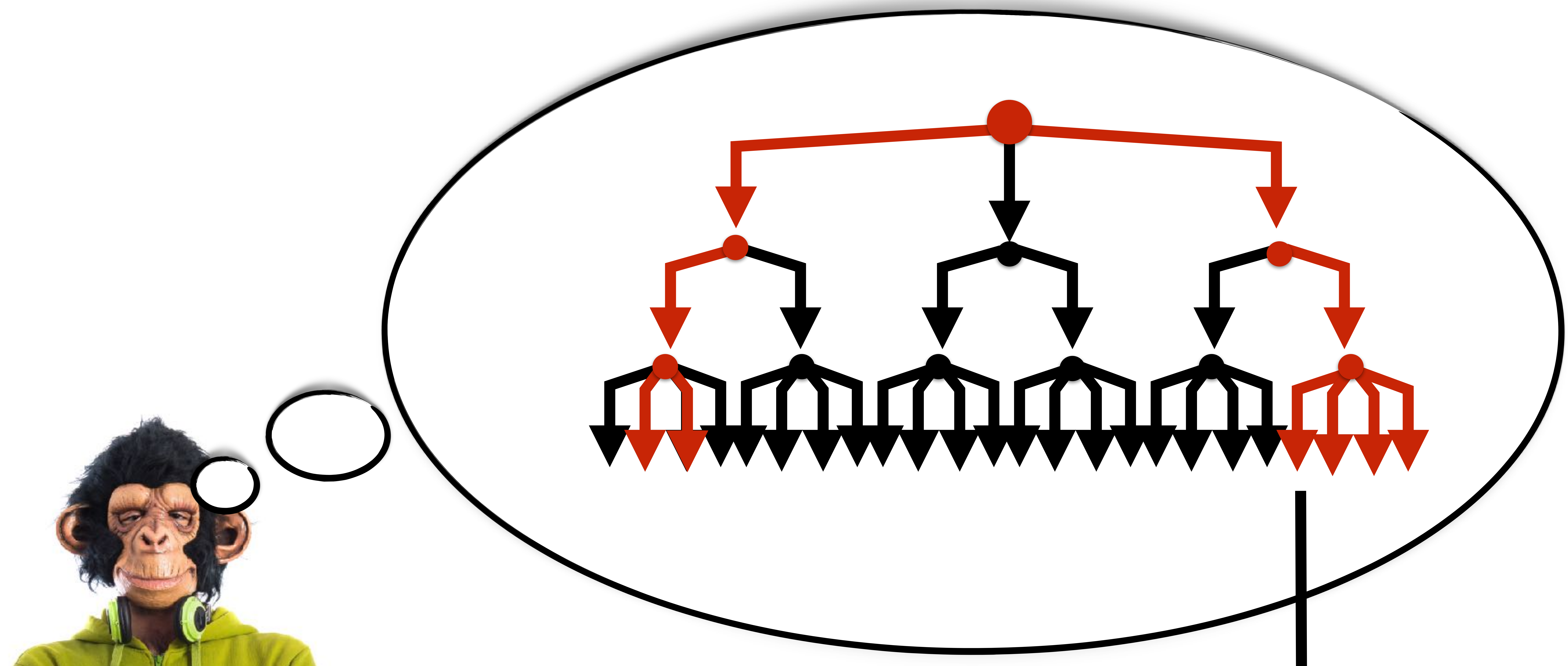

p = 0.
p = 0.7
p = 0.07



p = 0.
p = 0.7
p = 0.045
p = 0.11
p = 0.24
p = 0.07



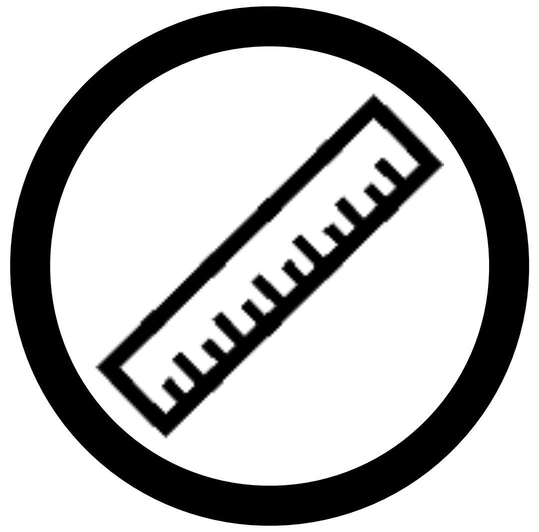

Report only
successful
pipeline



Report all pipelines but do not correct

Summary

Is there a relationship between **autism** and **vaccination**?



measure some stuff…

Evaluate that measured stuff using **statistical inference**

Publish findings in a scientific journal

things that can go wrong…

sampling error

can lead to wrong inferences about
the underlying population

dichotomous decision making

is subject to false positives and
false negatives

things that can go wrong…

sampling error

can lead to wrong inferences about the underlying population

analytical flexibility

can amplify human error and bias

dichotomous decision making

is subject to false positives and false negatives

the publication system

rewards certain results more than others

Full References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of verbal learning and verbal behavior, 12(4), 335-359.

Fox, N., Honeycutt, N., & Jussim, L. (2018). How Many Psychologists Use Questionable Research Practices? Estimating the Population Size of Current QRP Users.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science, 23(5), 524-532.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. Laboratory Phonology: Journal of the Association for Laboratory Phonology, 10(1).

Schönbrodt, F. D. (2016). p-hacker: Train your p-hacking skills! Retrieved from http://shinyapps.org/apps/p-hacker/.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science, 22(11), 1359-1366.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. Journal of Memory and Language, 103, 151-175.

Winter, B. (2011). Pseudoreplication in phonetic research. In Proceedings of the international congress of phonetic science: Hong Kong (pp. 2137-2140).

Winter, B. (2015). The other N: the role of repetitions and items in the design of phonetic experiments. In Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow: The University of Glasgow.