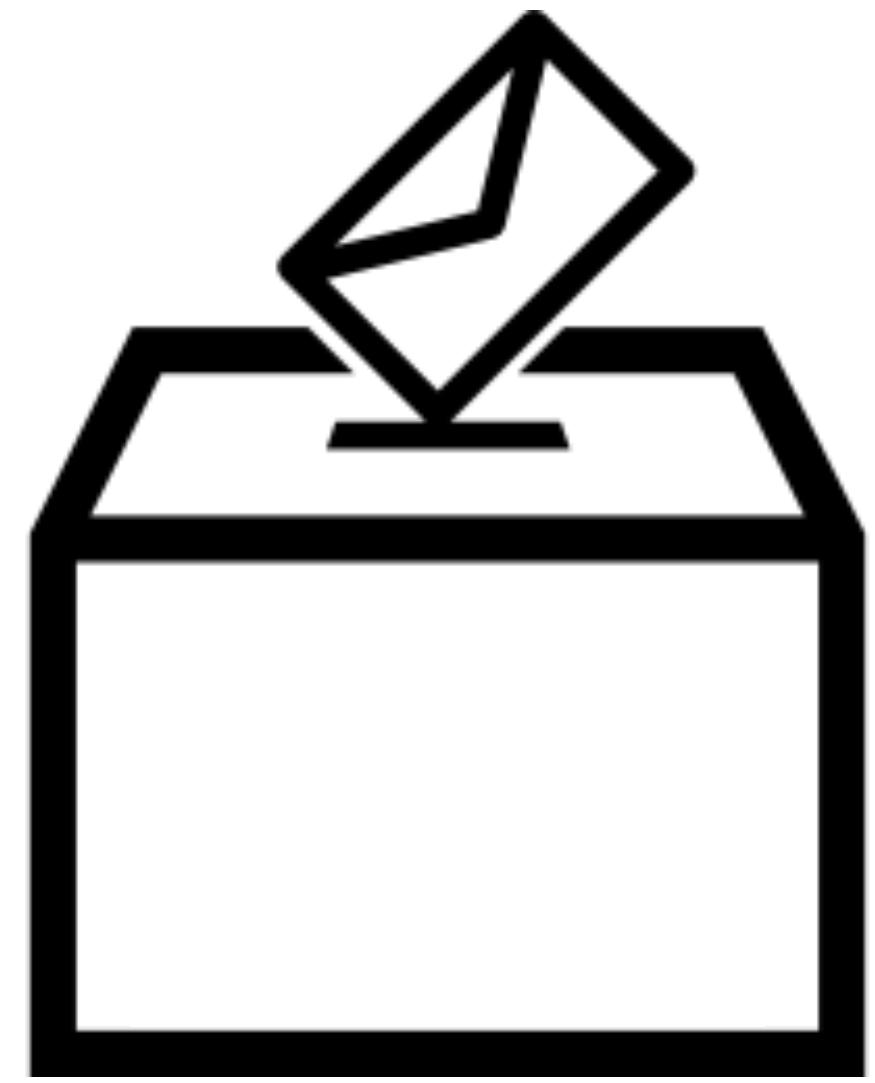


Statistical inference and things that can go wrong

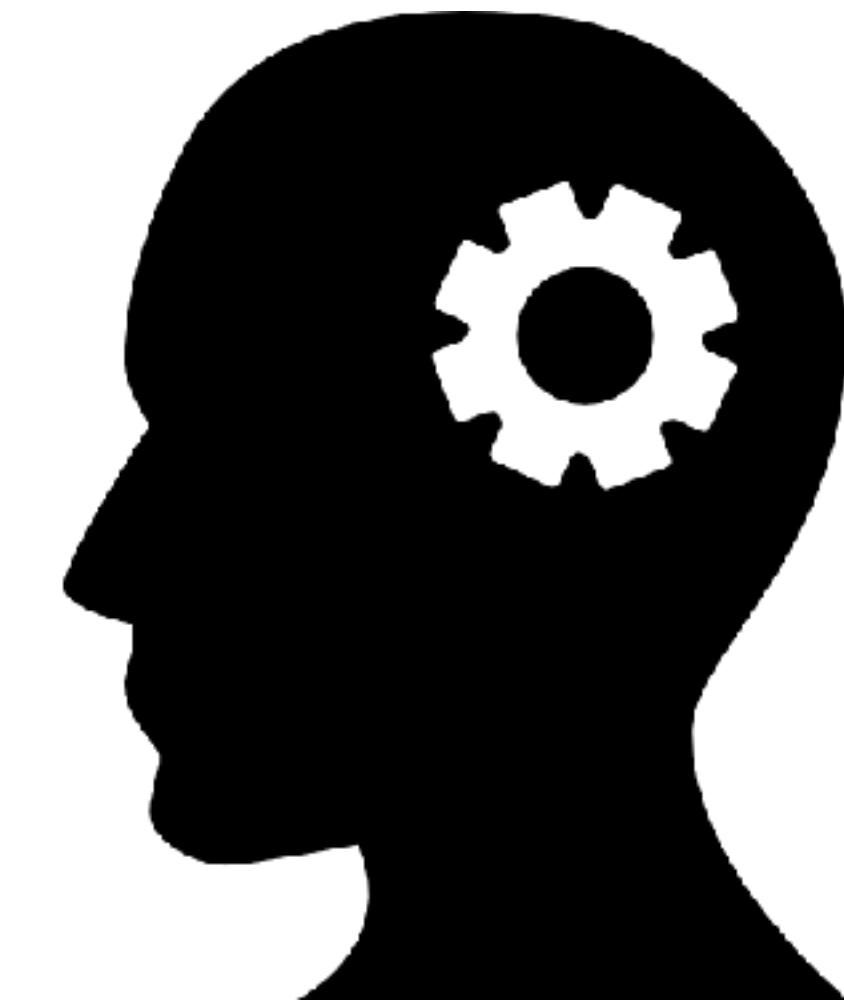
Timo B. Roettger



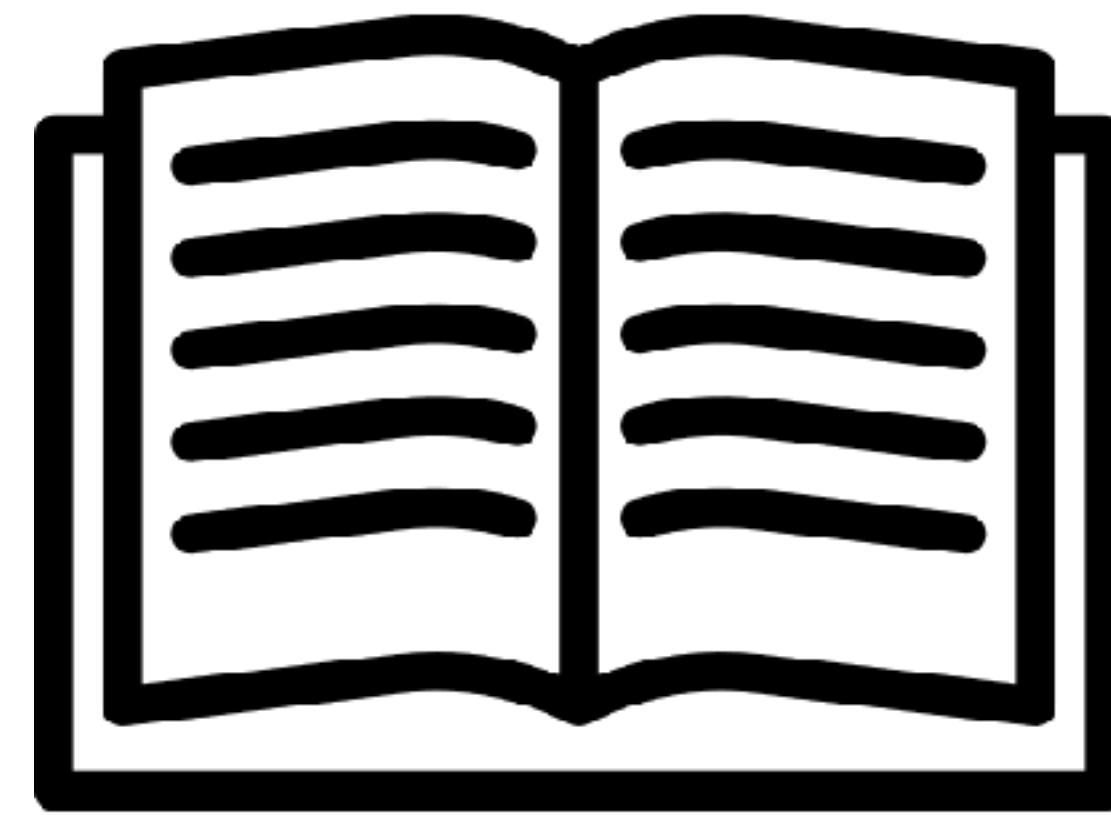
@TimoRoettger



**election
polls**



**human
cognition**



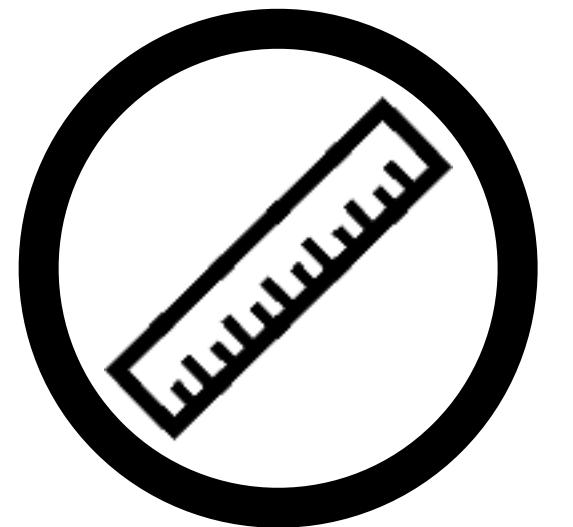
**scientific
publications**



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



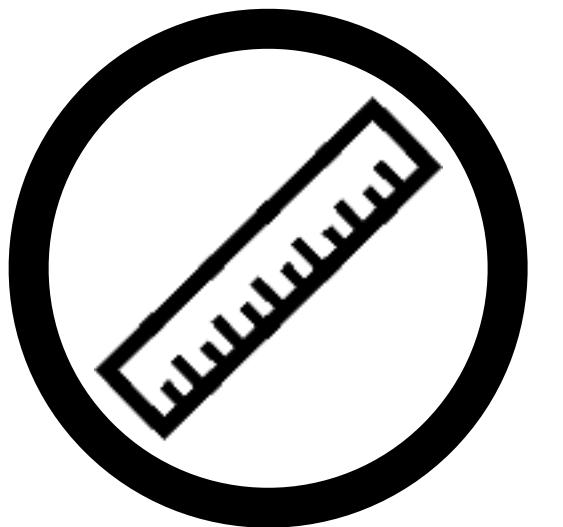
Is there a relationship
between **autism** and
vaccination?



measure some stuff...



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



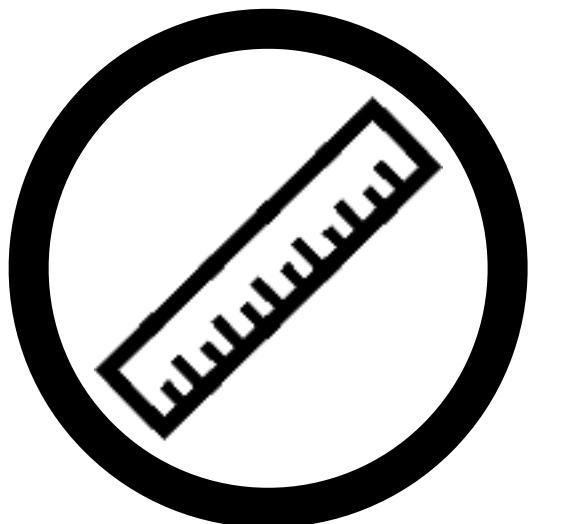
measure some stuff...



Evaluate that
measured stuff
using **statistical
inference**



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



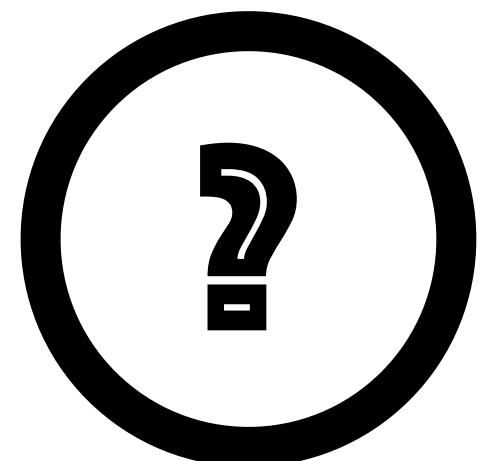
measure some stuff...



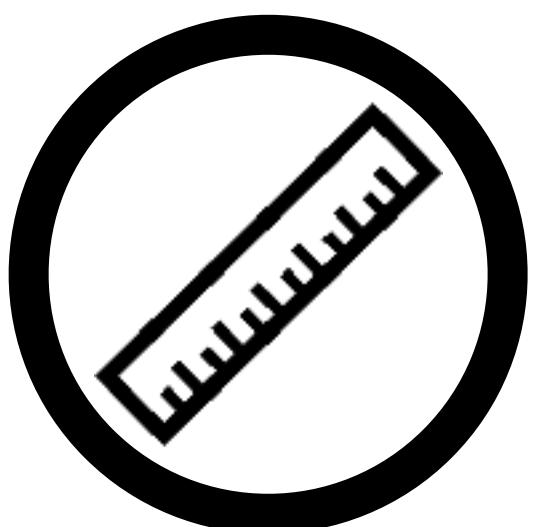
Evaluate that
measured stuff
using **statistical
inference**



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



Donald J. Trump 
@realDonaldTrump

Yuge scientific finding: Vaccination caused autism in billions of children.

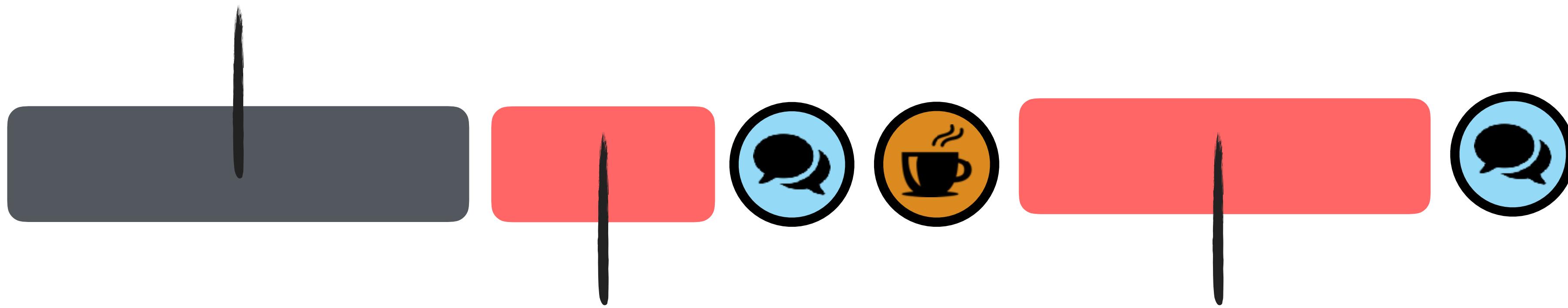
RETWEETS LIKES
7,423 **15,172**

2:09 PM - 13 Jan 2020

912 7K 15K



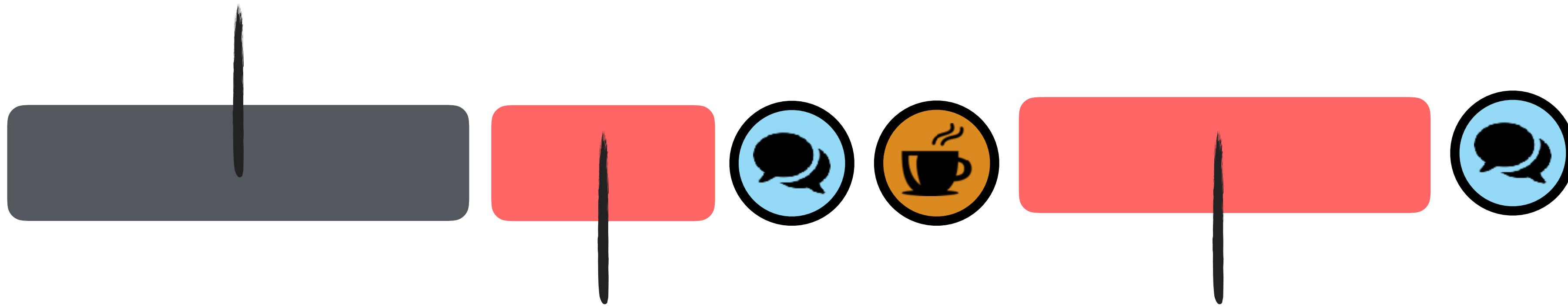
Statistical inference and NHST



the **infamous**
p-value

Things that
can go
terribly **wrong**

Statistical inference and NHST



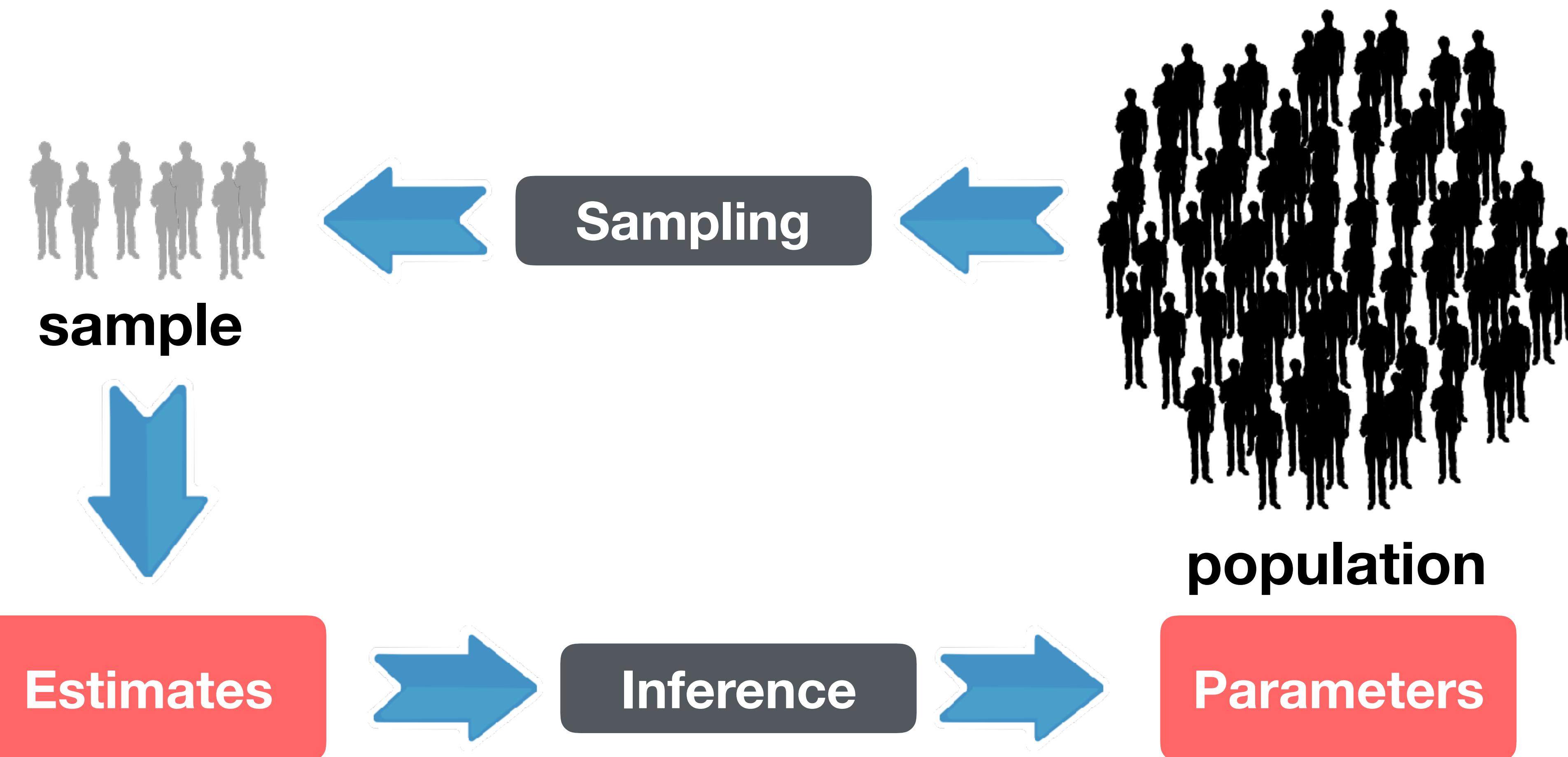
the **infamous**
p-value

Things that
can go
terribly **wrong**

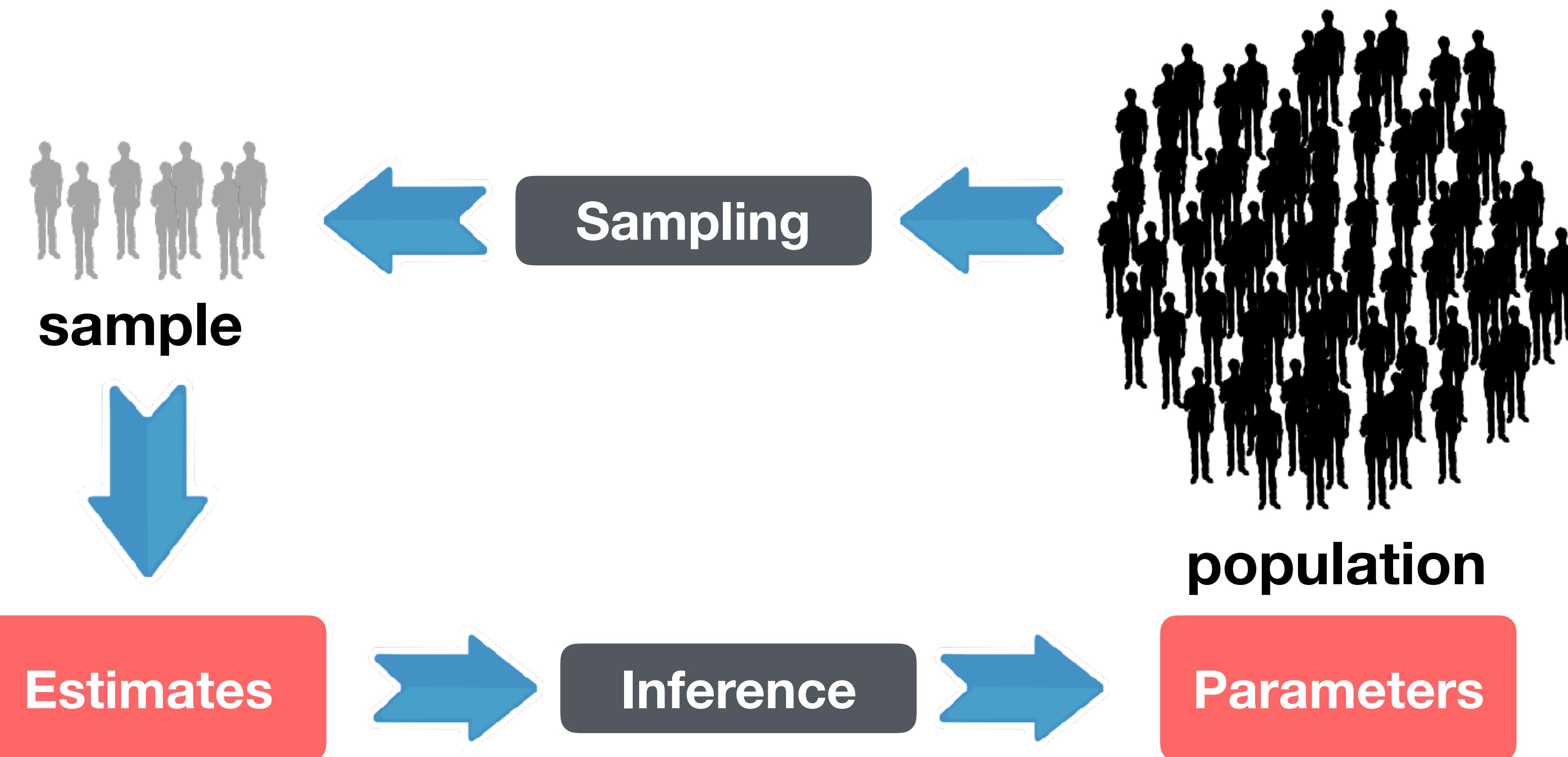
Statistical inference



Statistical inference



Statistical inference



NHST

Null-Hypothesis Significance Testing

NHST

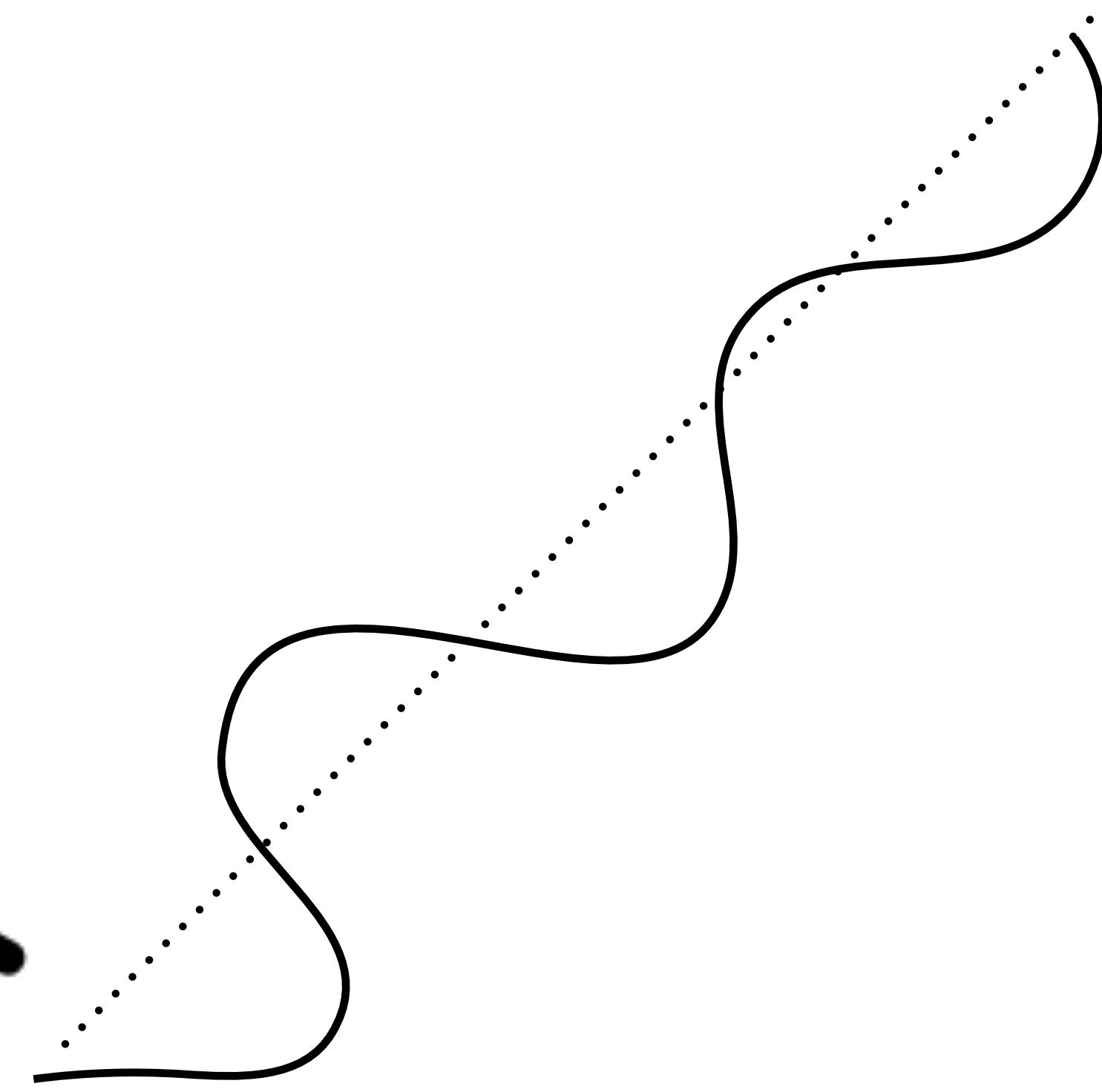
Null-Hypothesis Significance Testing

0. Set up the Alternative Hypothesis (H_a).
1. Set up the Null-Hypothesis (H_0).



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.

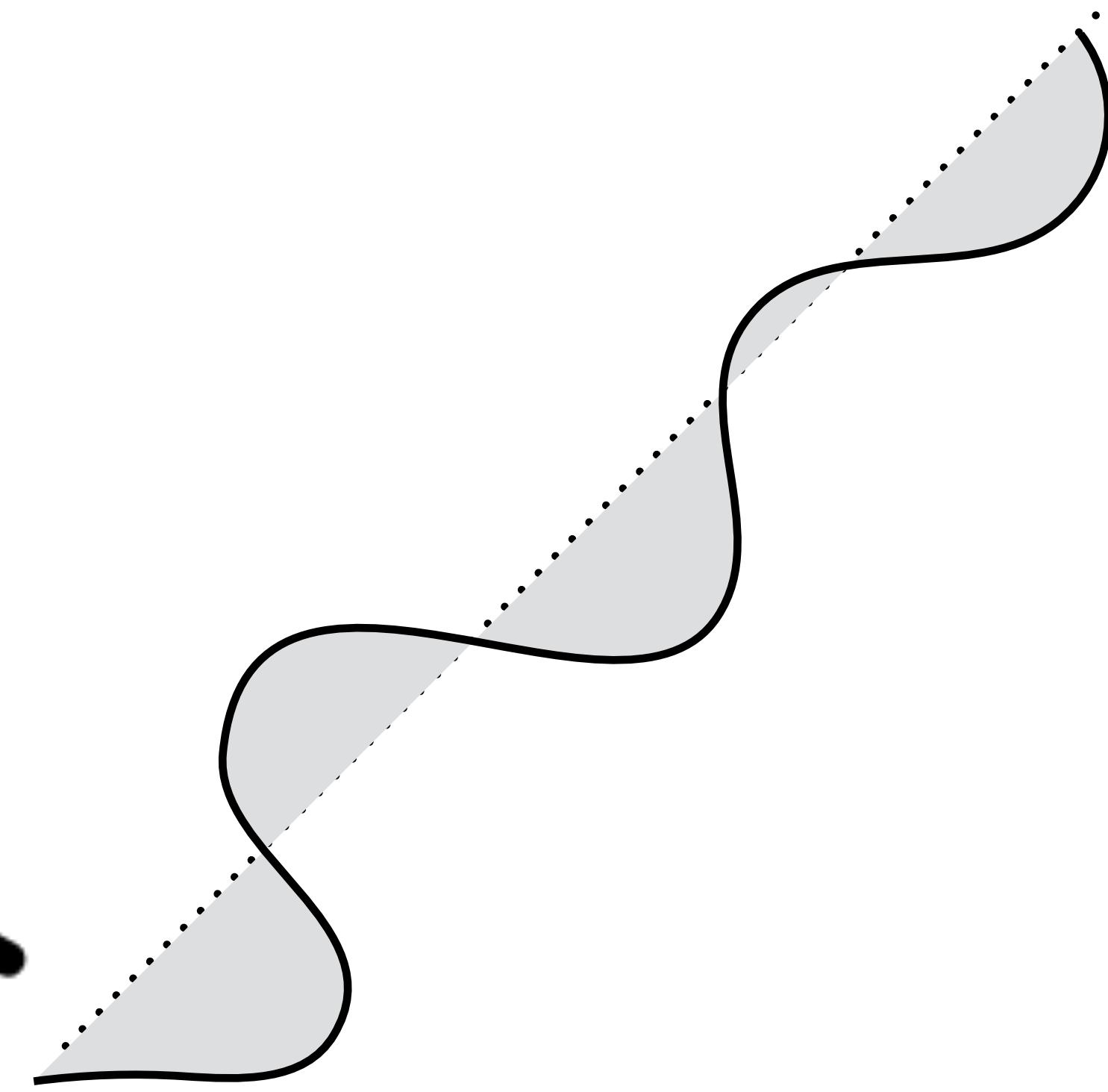


H_a : Drunken people diverge more from a straight line than sober people.

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

difference between groups

* 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

the larger *t*,
the smaller *p*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{SD^2}{N}}}$$

standard error (SE)

“noise”

sample size

Measuring the compatibility of the data

difference between groups

the larger t ,
the smaller p

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{SD^2}{N}}}$$

standard error (SE)

“noise”

sample size

Measuring the compatibility of the data

difference between groups

t value

9.8

$$= \frac{4 - 2}{\sqrt{\frac{1}{24}}}$$

standard error (SE)

“noise”

sample
size

Measuring the compatibility of the data

difference between groups

$$t \text{ value} = \frac{4 - 3.5}{\sqrt{\frac{1}{24}}}$$

standard error (SE)

“noise”

sample size

The diagram illustrates the formula for a t-value. The t-value is shown as 2.5. The formula is $t \text{ value} = \frac{4 - 3.5}{\sqrt{\frac{1}{24}}}$. A blue arrow points down to the t-value. A green brace groups the standard error (SE) term, which is $\sqrt{\frac{1}{24}}$. A red arrow points down to the sample size term, which is 24. The term $4 - 3.5$ is labeled "difference between groups". The term $\frac{1}{24}$ is labeled "noise".

Measuring the compatibility of the data

difference between groups

$$t \text{ value} = \frac{\text{difference between groups}}{\sqrt{\frac{\text{"noise"}}{\text{sample size}}}}$$

↑ **5.5** = **4 - 3.5**

standard error (SE) { **0.2** / **24** }

“noise” ↓
sample size

The diagram illustrates the formula for calculating a t-value. The t-value is shown as a blue box containing '5.5'. Above it, the text 't value' is written in blue. To the right of the t-value, the formula is displayed: $t \text{ value} = \frac{\text{difference between groups}}{\sqrt{\frac{\text{"noise"}}{\text{sample size}}}}$. The 'difference between groups' is shown in a pink box as '4 - 3.5'. The denominator is enclosed in a black square root symbol. Inside the square root, the 'noise' is shown in a light green box as '0.2' above a line, and the 'sample size' is shown in a light green box as '24' below a line. A green curly brace groups the 'noise' and 'sample size' boxes. A blue arrow points upwards from the text 't value' towards the t-value itself. A green arrow points downwards from the text 'sample size' towards the 'sample size' box.

Measuring the compatibility of the data

difference between groups

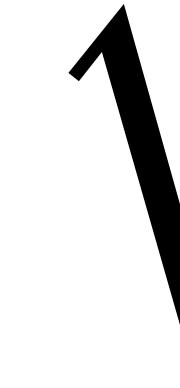
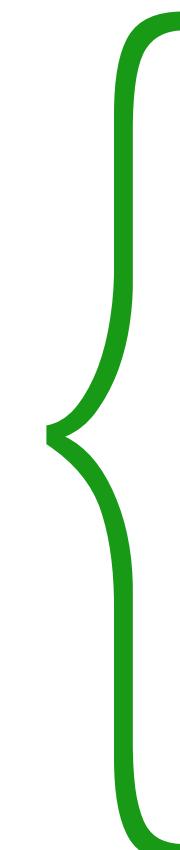
t value



2.7

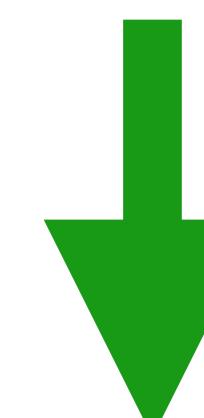
$$= \frac{4 - 3.5}{\sqrt{\frac{0.2}{6}}}$$

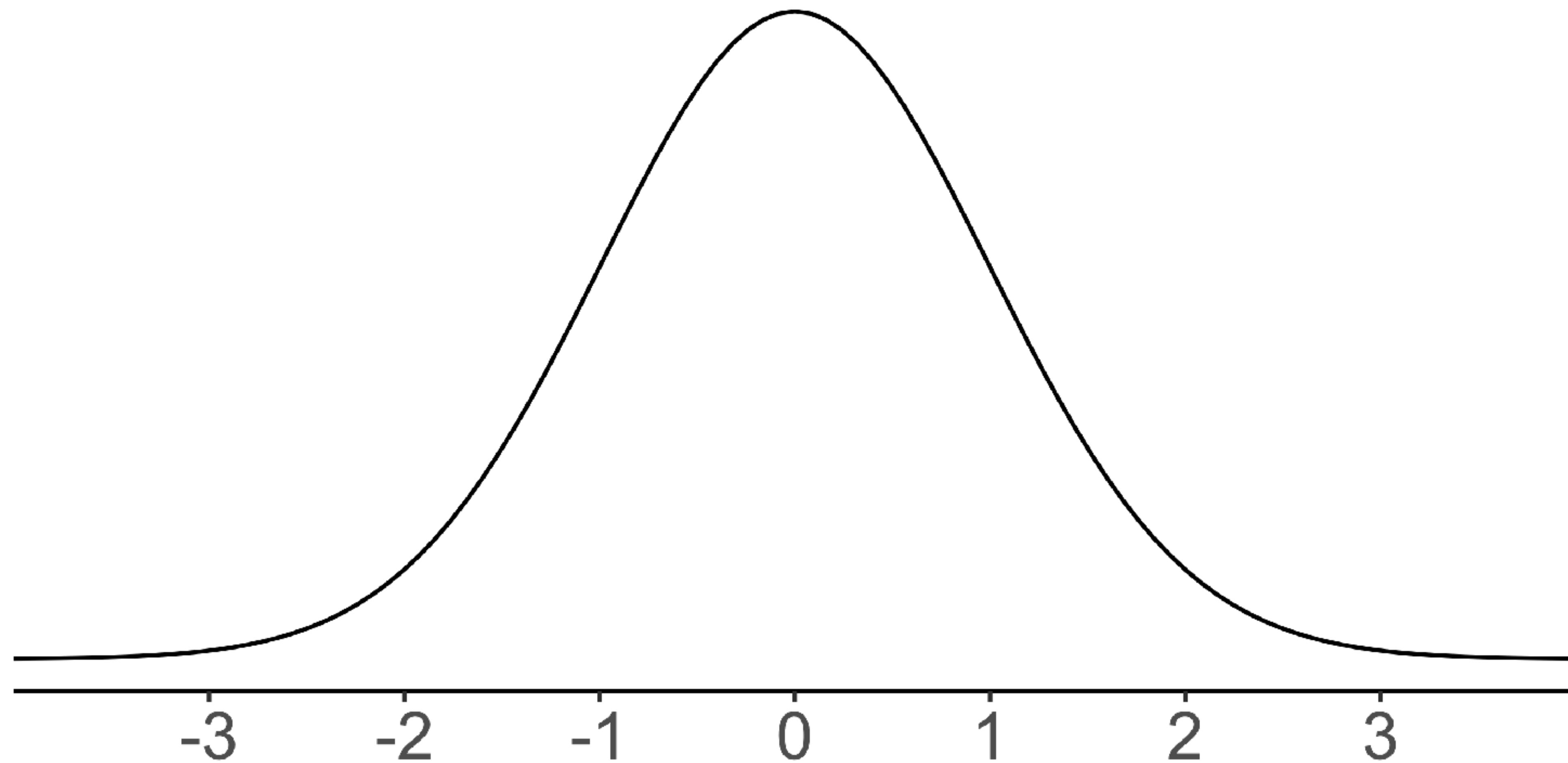
standard error (SE)



$$\sqrt{\frac{0.2}{6}}$$

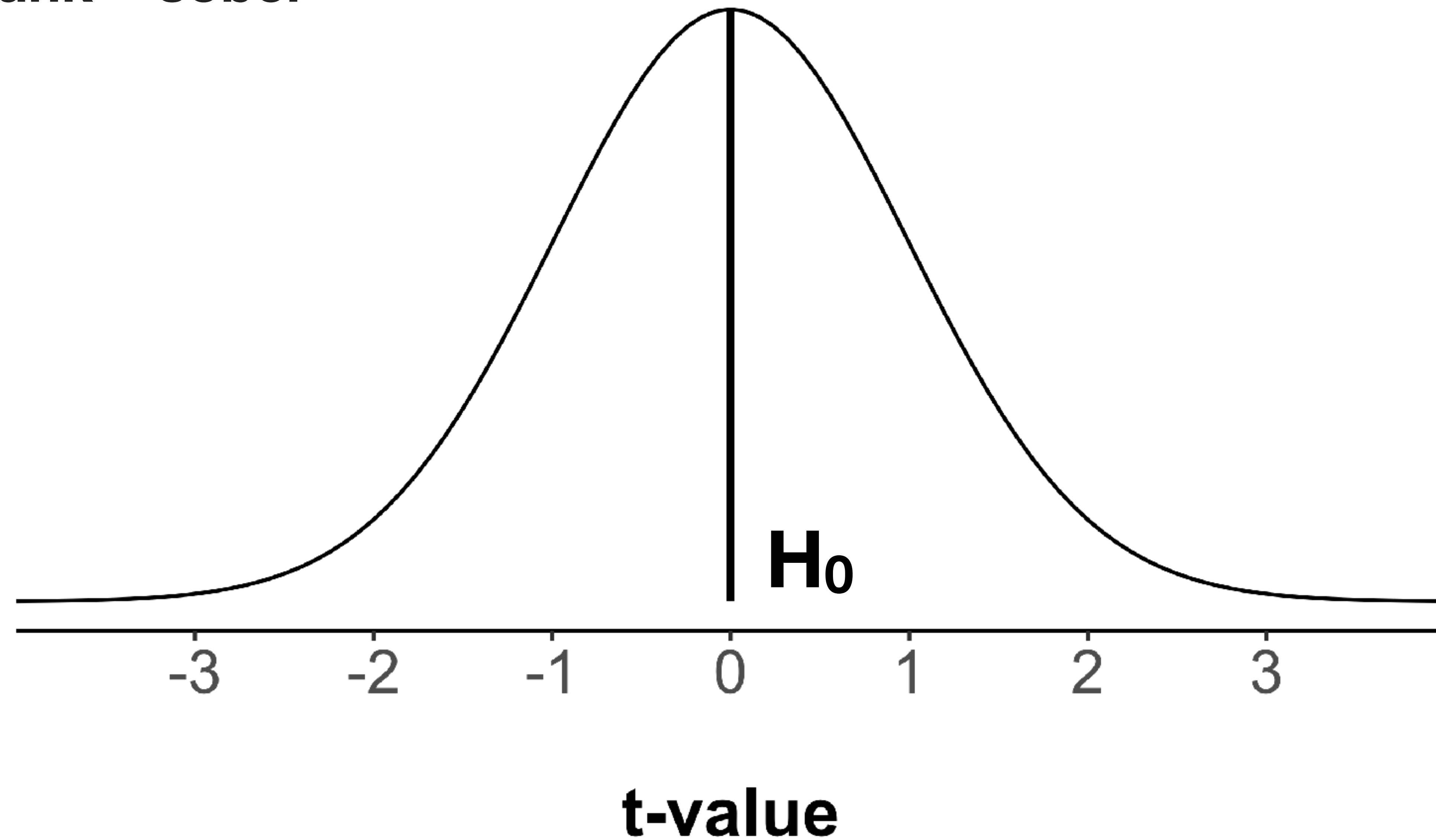
“noise”
sample
size



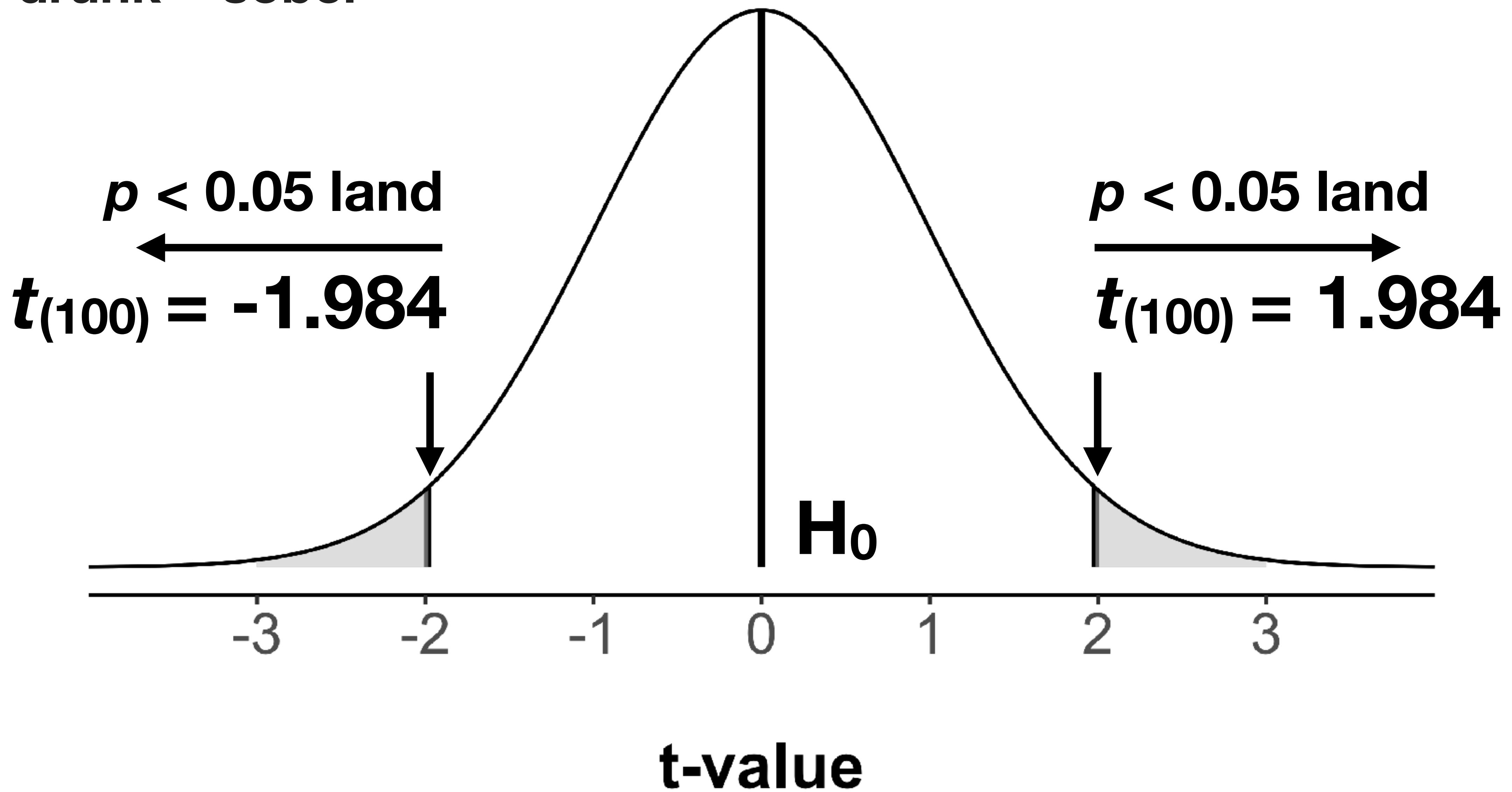


t-value

H_0 : drunk = sober

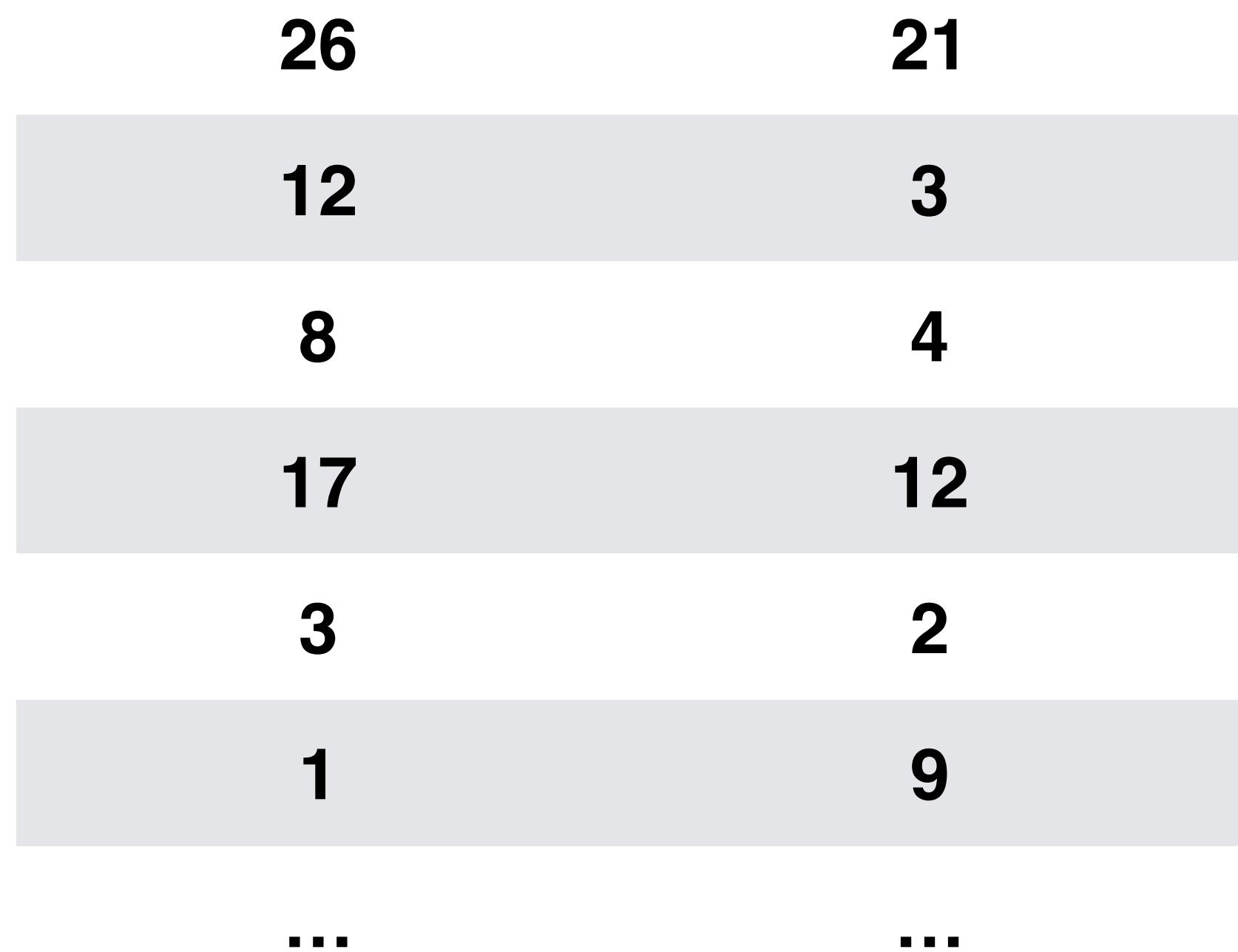


H_0 : drunk = sober





drunken sober

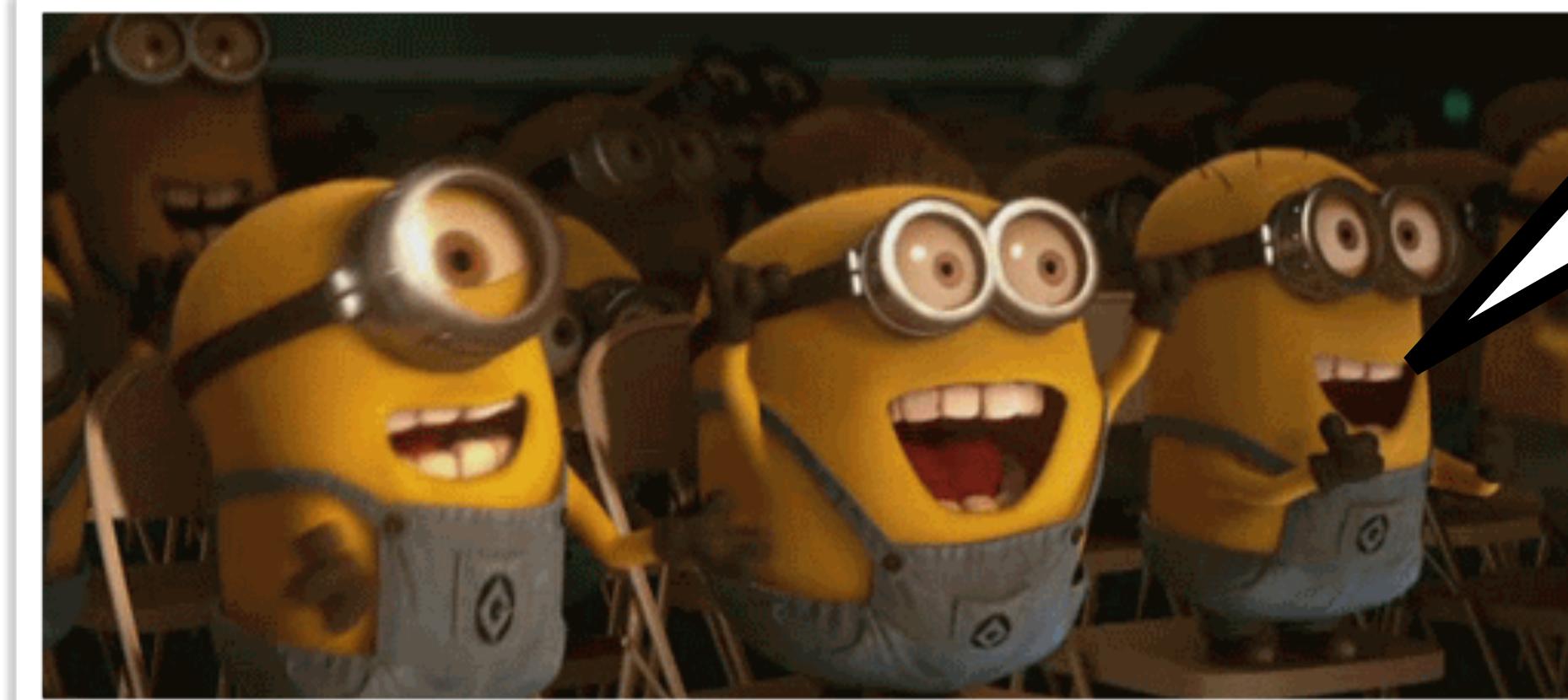


$$2.6 = \frac{6.6 - 5.6}{2.9}$$

A mathematical equation showing the calculation of a difference. The result 2.6 is in a blue box. The numerator 6.6 - 5.6 is highlighted in a pink box. The denominator 2.9 is in a green box. A large black arrow points from the pink box down to the green box.

20

$H_0: \text{drunk} = \text{sober}$



$H_a: \text{drunk} > \text{sober}^*$

significant

$$p = 0.00524$$

$$t_{(100)} = 2.6$$

H_0



t-value



You are DRUNK.



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.

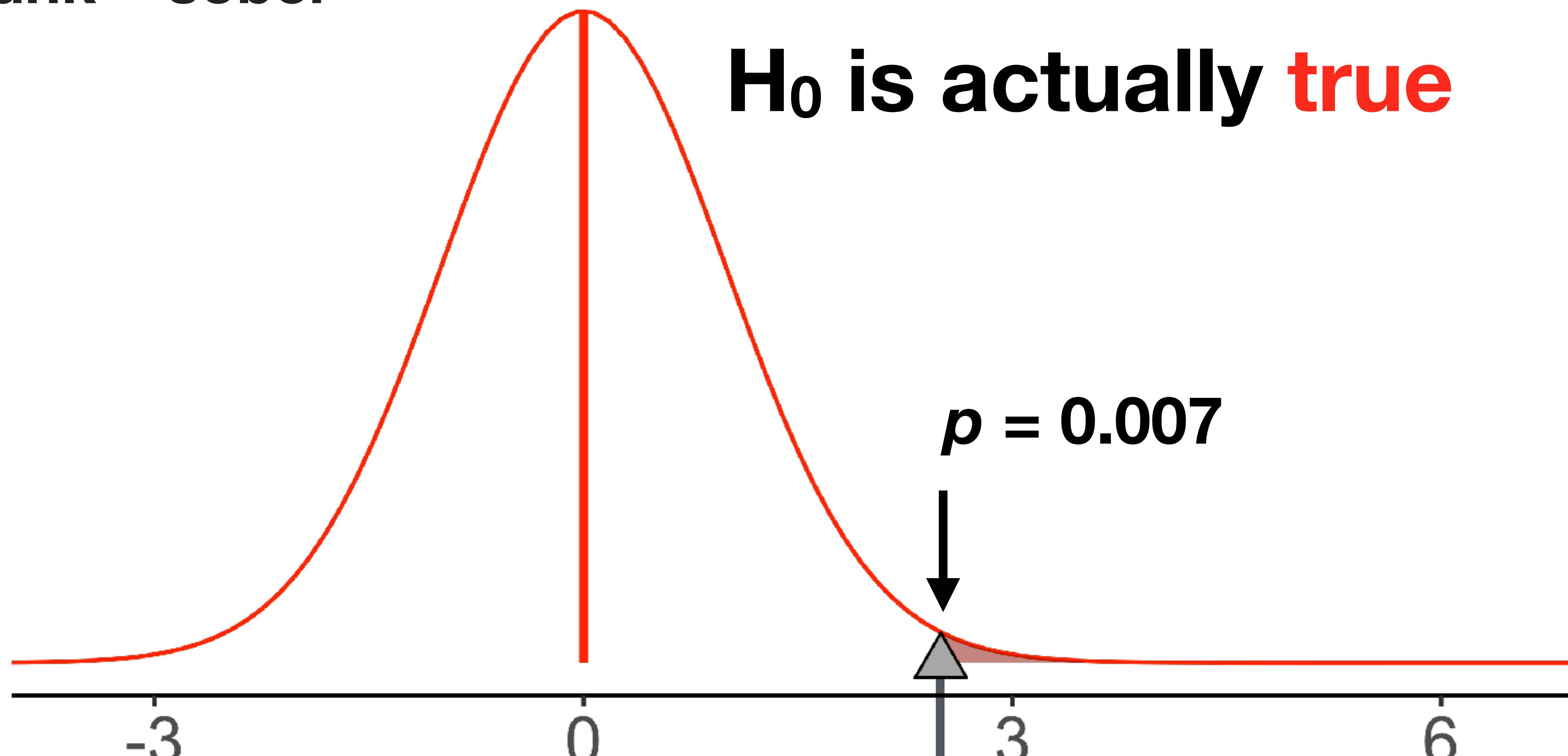
You are **DRUNK.**

False Positive Type-I error



H_0 : drunk = sober

H_0 is actually **true**



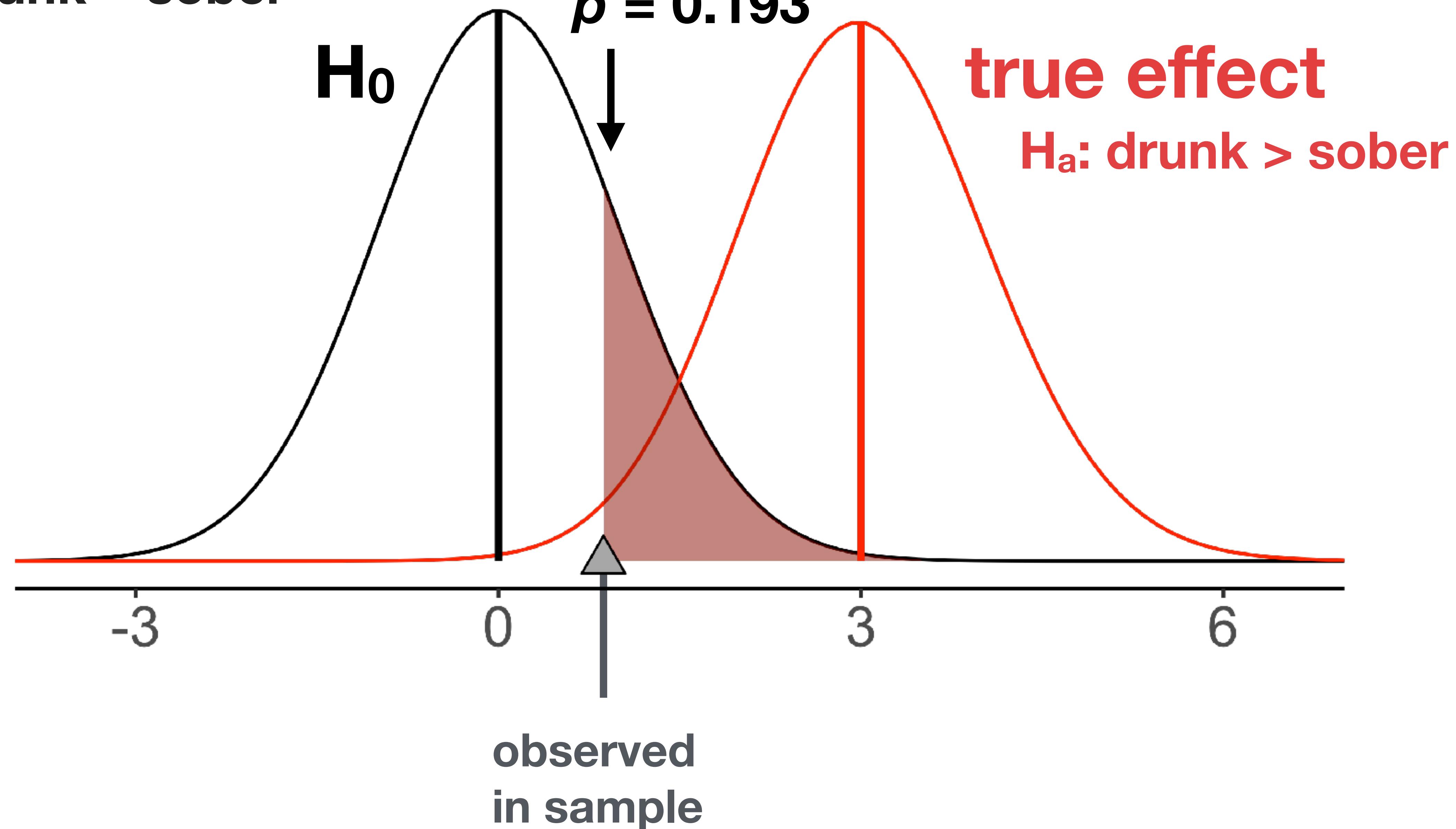
observed
in sample



You are **NOT** drunk.

False Negative Type-II error

H_0 : drunk = sober



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

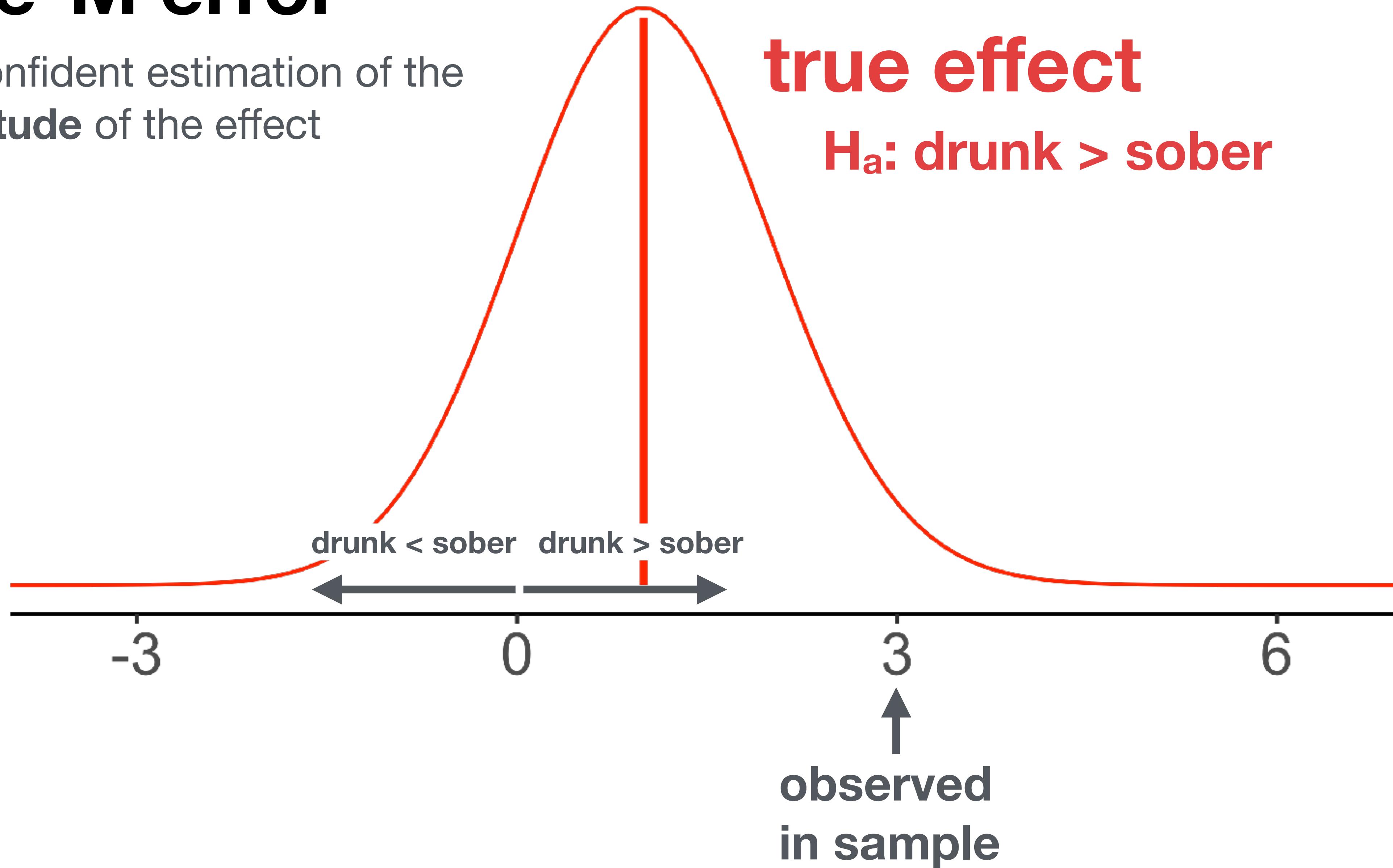
Overconfident estimation of the **magnitude** of the effect

Type-S error

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

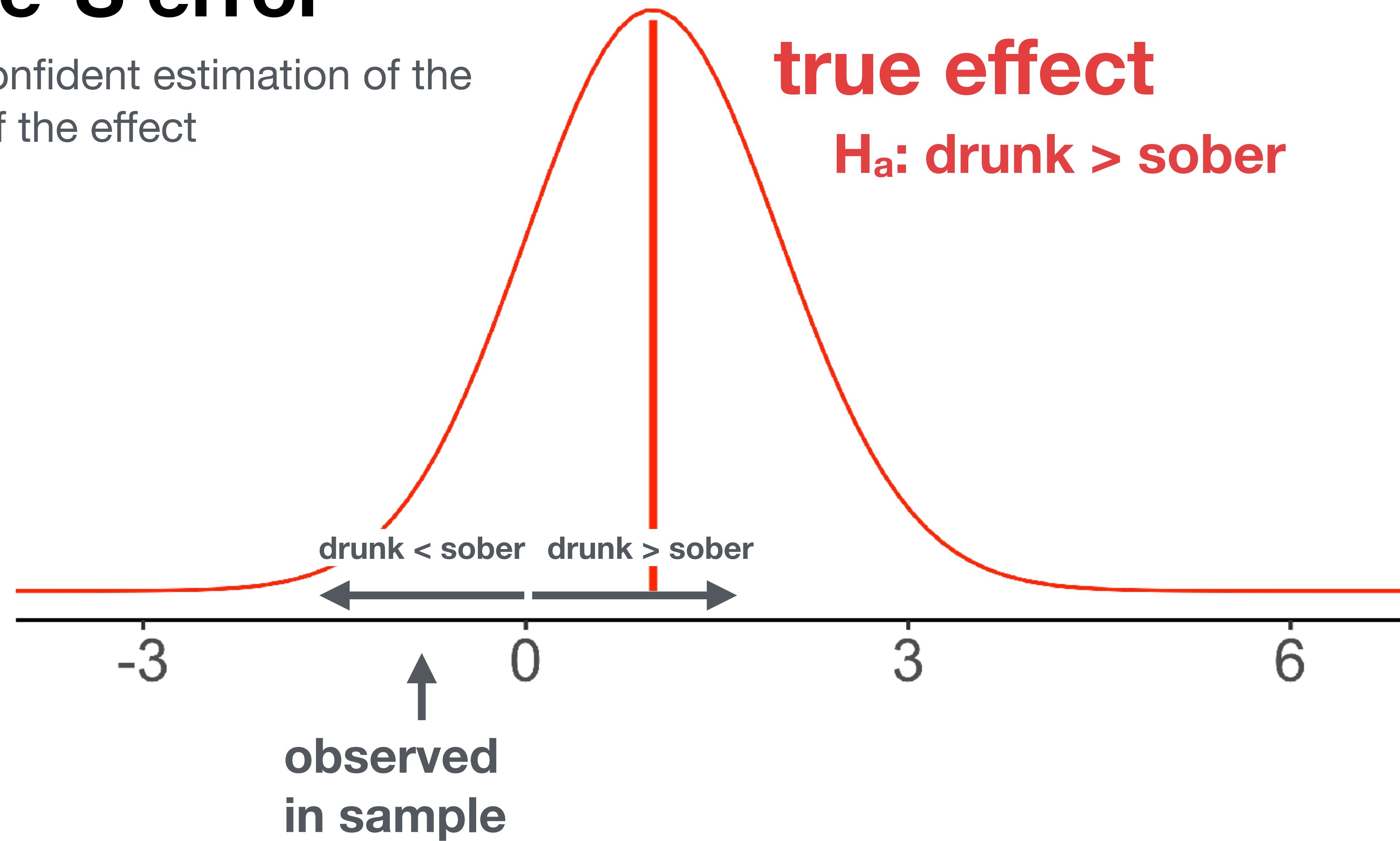
Type-M error

Overconfident estimation of the
magnitude of the effect



Type-S error

Overconfident estimation of the sign of the effect



$p < 0.05$



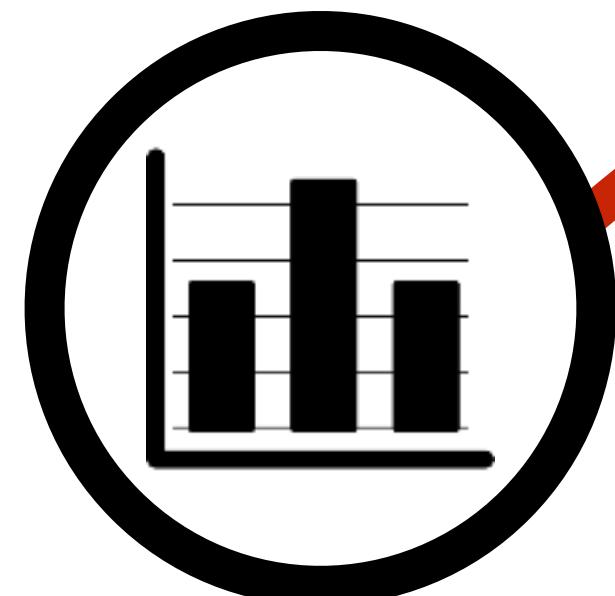
A photograph of two men cheering while riding a roller coaster. They are both shouting with their mouths wide open and have their arms raised. The man on the left is wearing glasses and a light-colored shirt. The man on the right is pointing his finger towards the camera. The background shows the steel structure and orange supports of the roller coaster. A black rectangular overlay covers the bottom left corner of the image.

$p < 0.05$

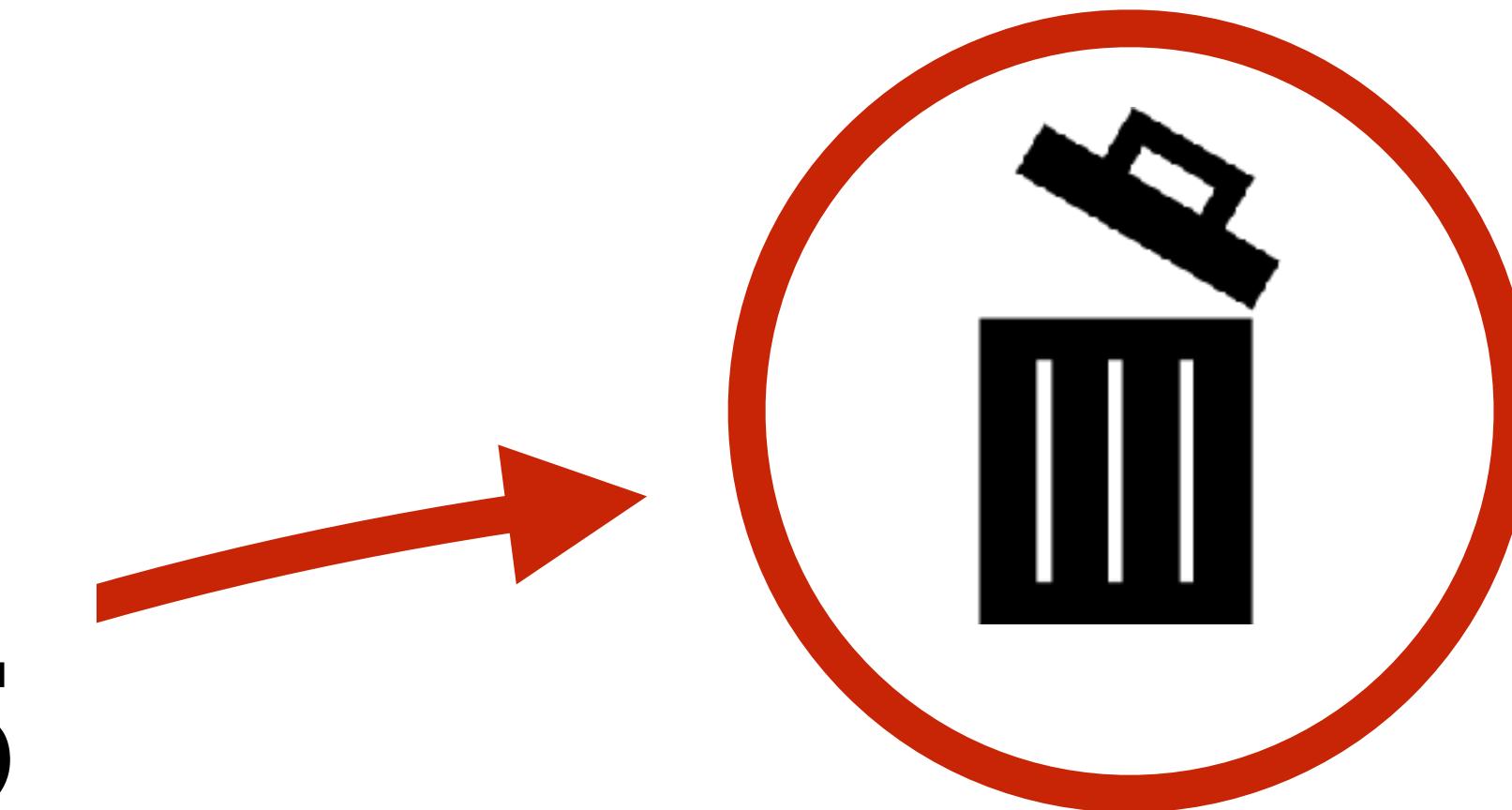
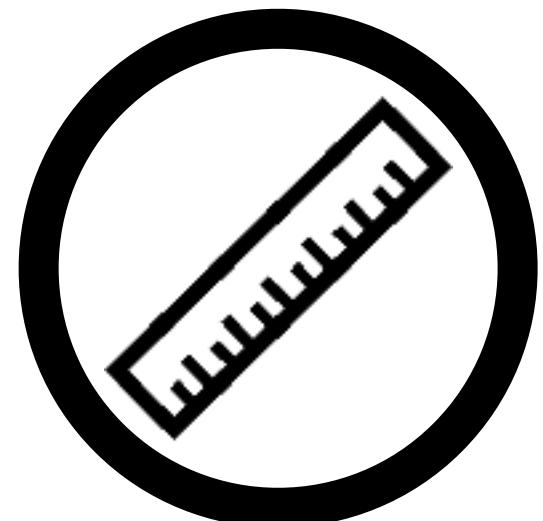
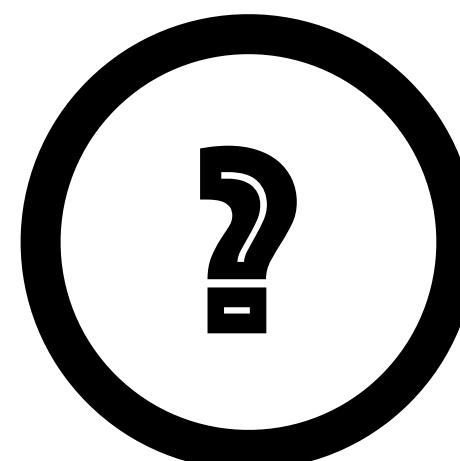


Evaluate that measured stuff using **statistical inference**

Publish findings in a scientific journal



not significant
 $p > 0.05$



Not published and never seen again



THE *p*-value

p-value

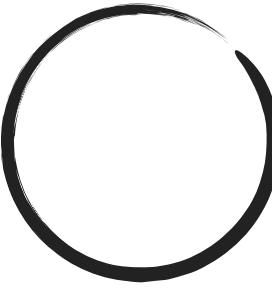
The probability of the **observed result**, plus more extreme results, **if the null hypothesis was true**



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**



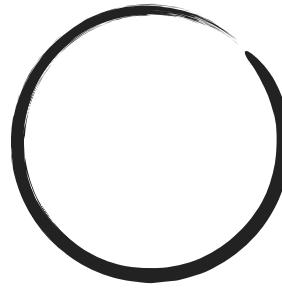
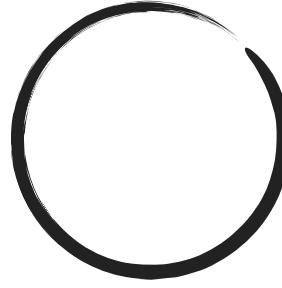
The probability of the null hypothesis



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



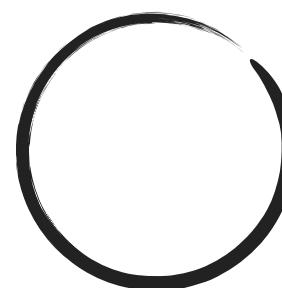
The probability of the alternative hypothesis



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



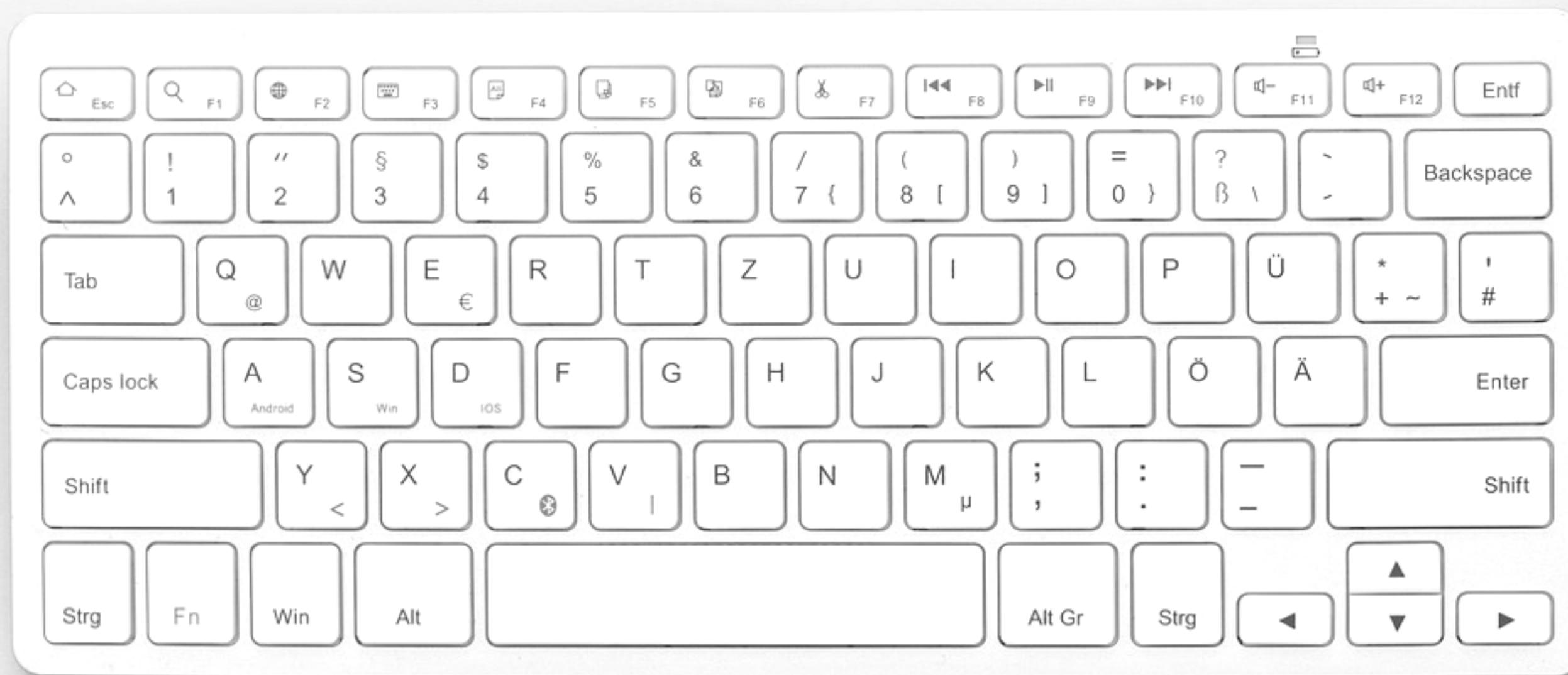
If < 0.05 , we can conclusively answer a scientific question





Unlucky sampling

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





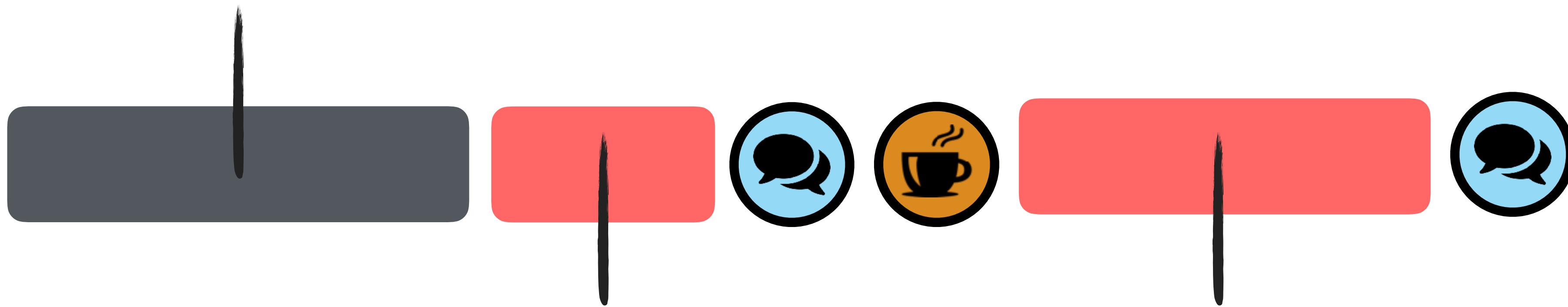
What are your questions?





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

$$t \text{ value} = \frac{4 - 2}{\sqrt{\frac{6}{24}}}$$

t value

standard error (SE)

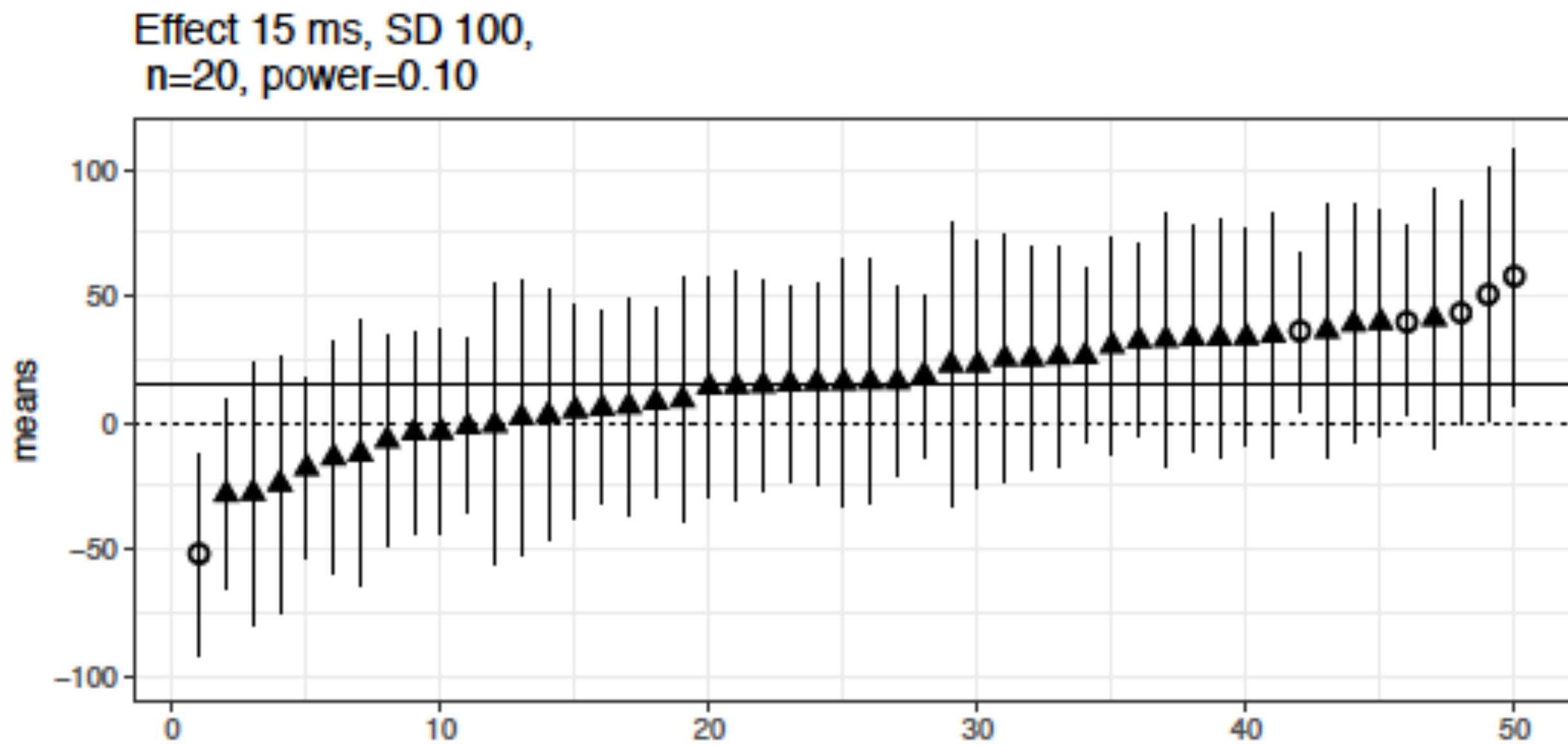
difference between groups

“noise”

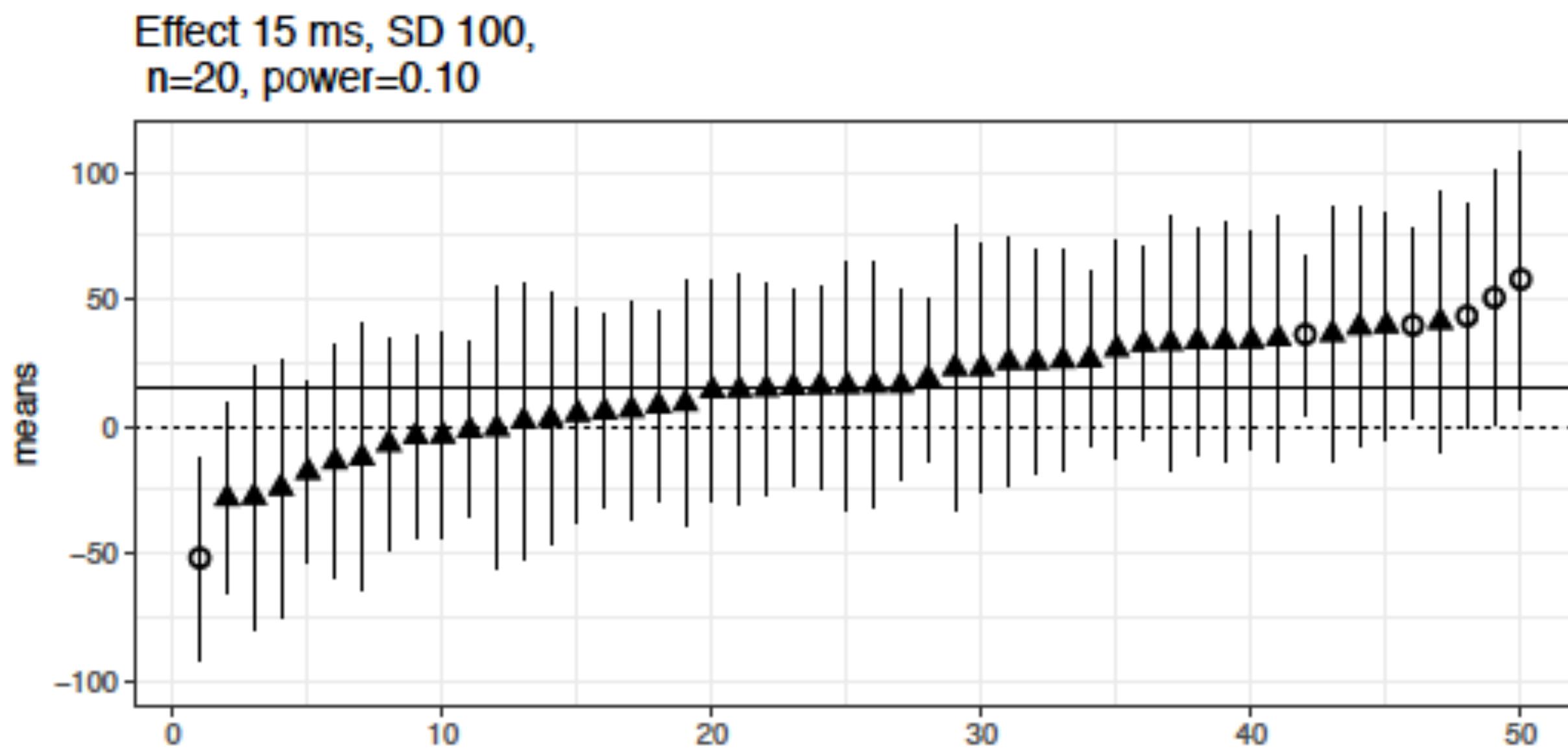
sample size

A diagram illustrating the formula for the t-value. The formula is $t \text{ value} = \frac{4 - 2}{\sqrt{\frac{6}{24}}}$. The term $4 - 2$ is highlighted in a pink box and labeled "difference between groups". The term $\sqrt{\frac{6}{24}}$ is highlighted in a light green box and labeled "noise". The number 24 is also highlighted in a light green box and labeled "sample size". A blue arrow points to the t-value. Brackets group the difference and standard error, and a green arrow points to the sample size.

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

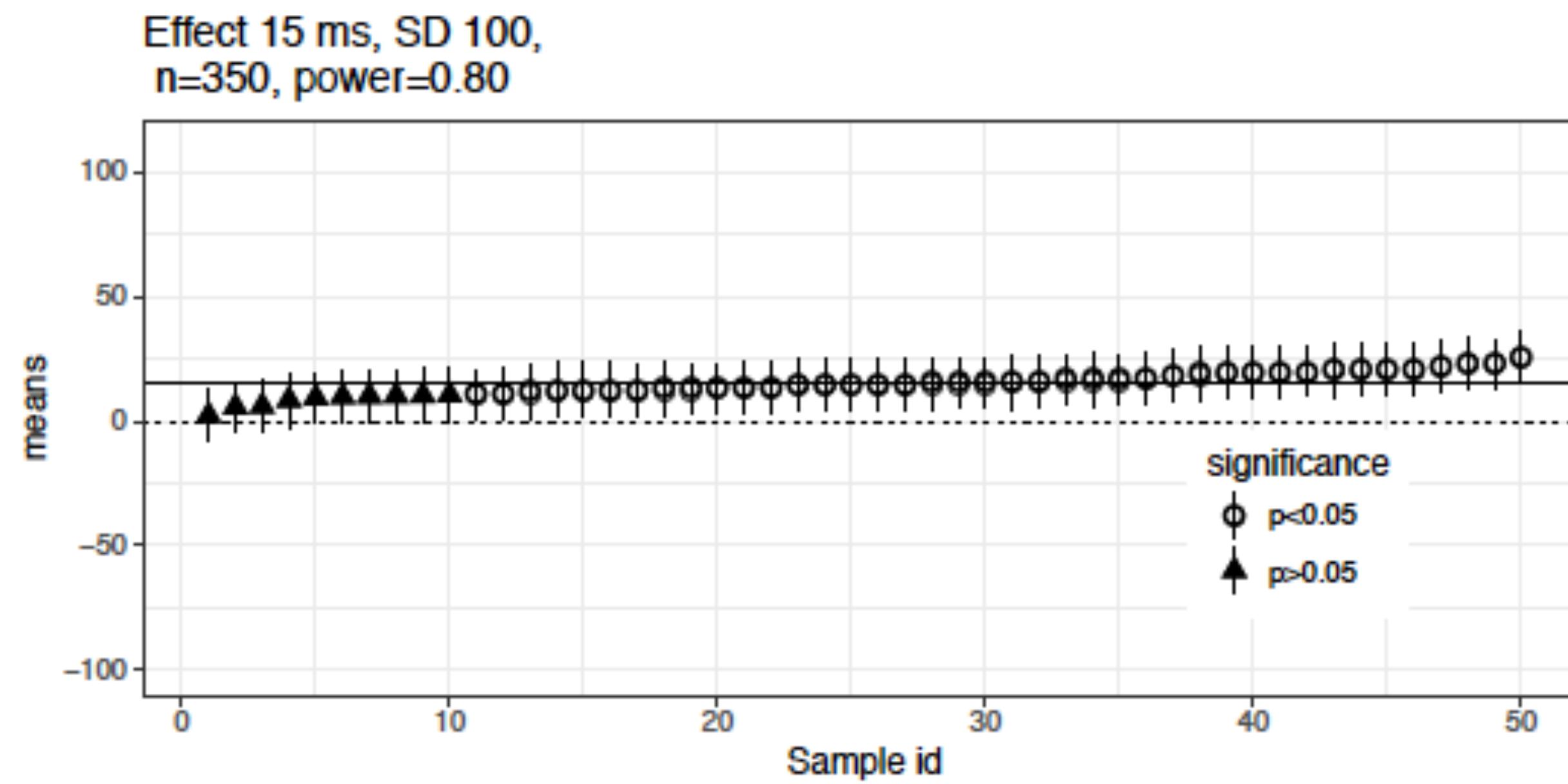
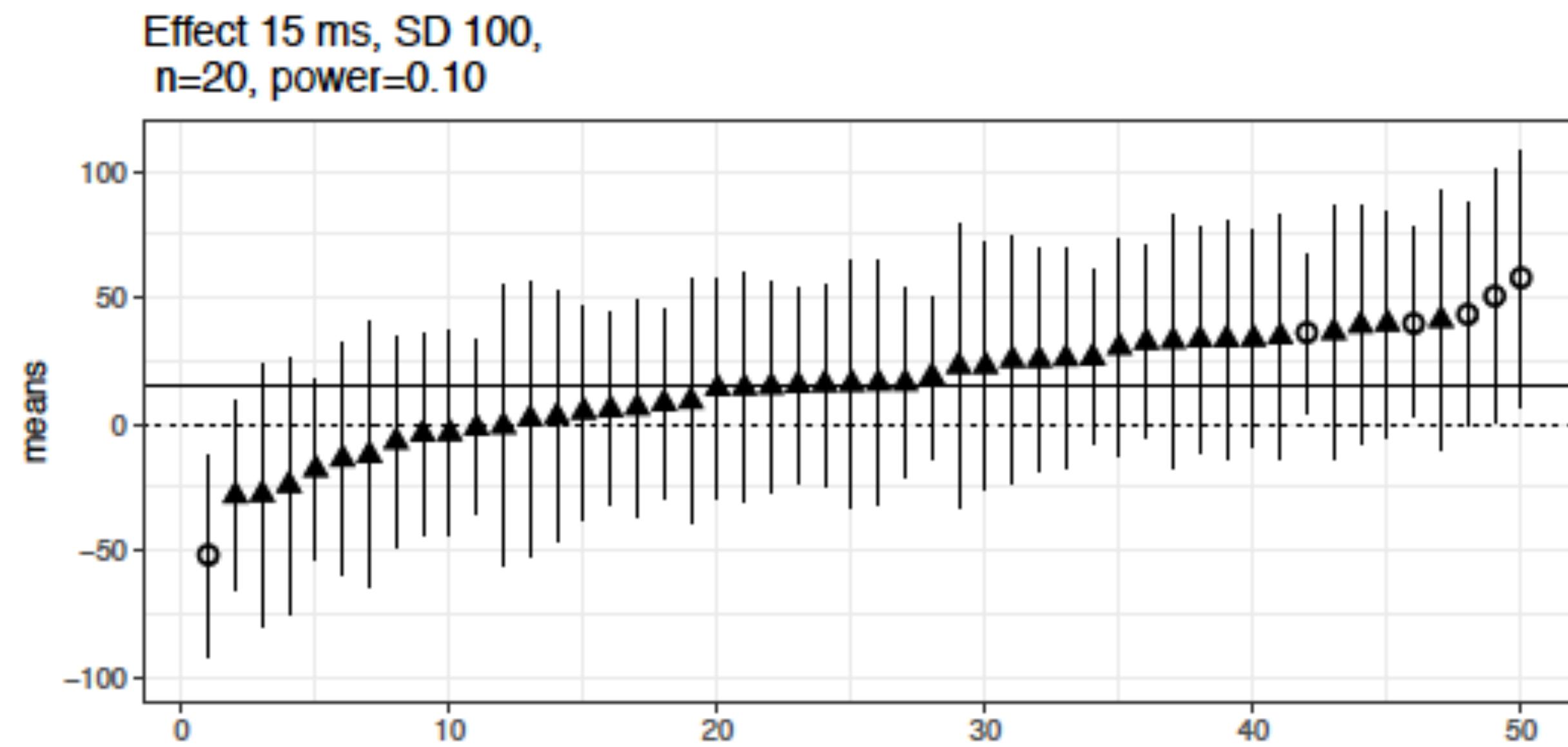


**Small
samples size
leads to increased
Type M &
Type S errors**



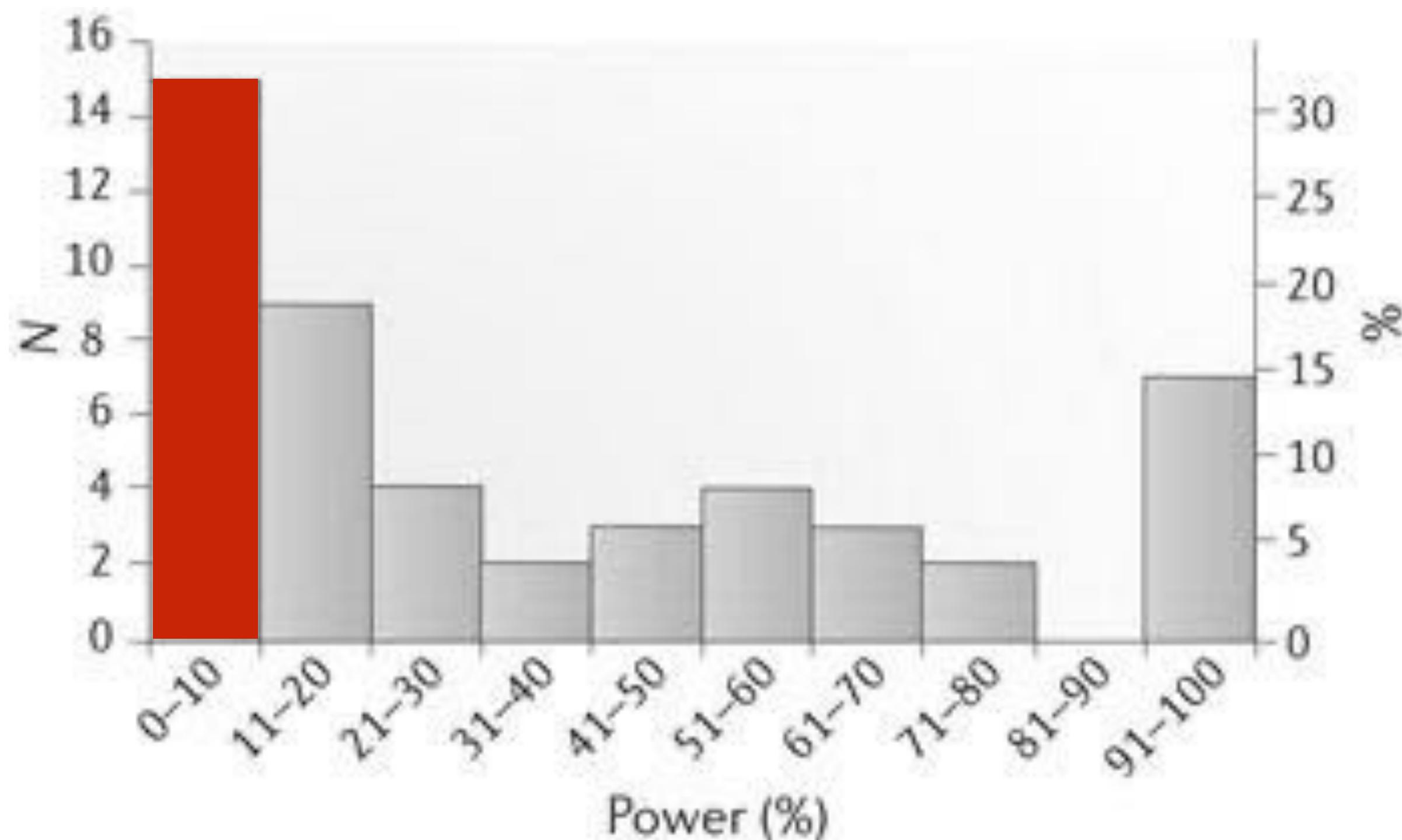
Effect 15 ms, SD 100,
 $n=350$, power=0.80

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



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

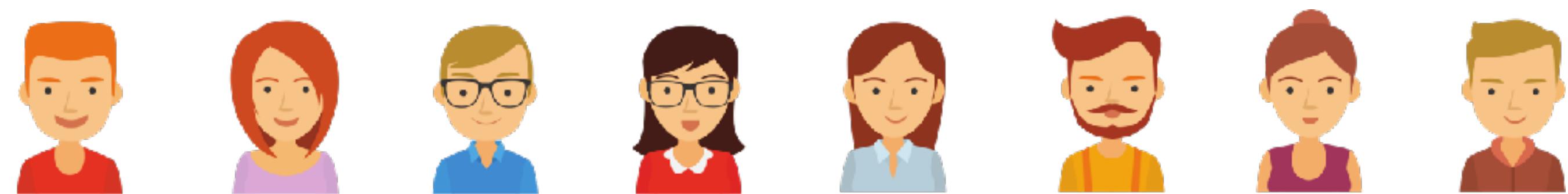
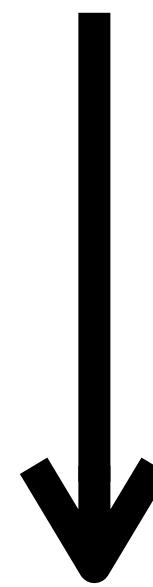


Derren Brown

Multiple testing

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

H_0 : Berlin = Osnabrück



More V-necks

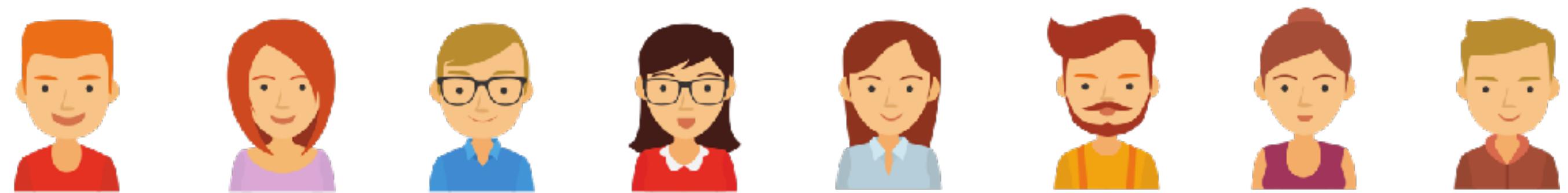
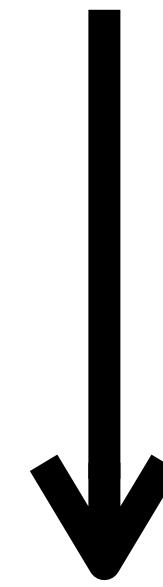
$p = 0.6$



Multiple testing

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

H_0 : Berlin = Osnabrück



More V-necks

$p = 0.6$

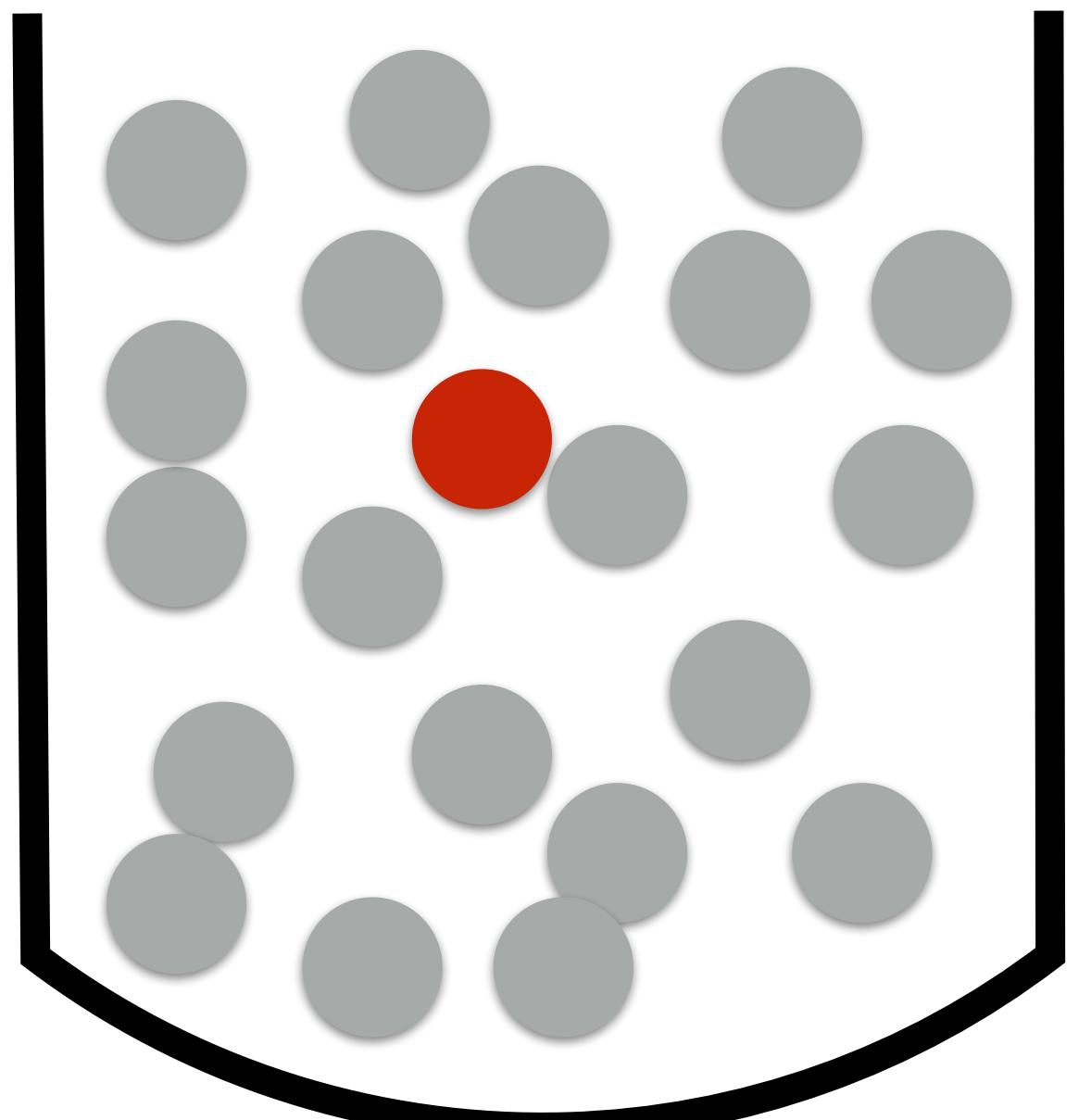


More colorful

$p = 0.04$

Probability of randomly pulling the **red** marble?

0.05



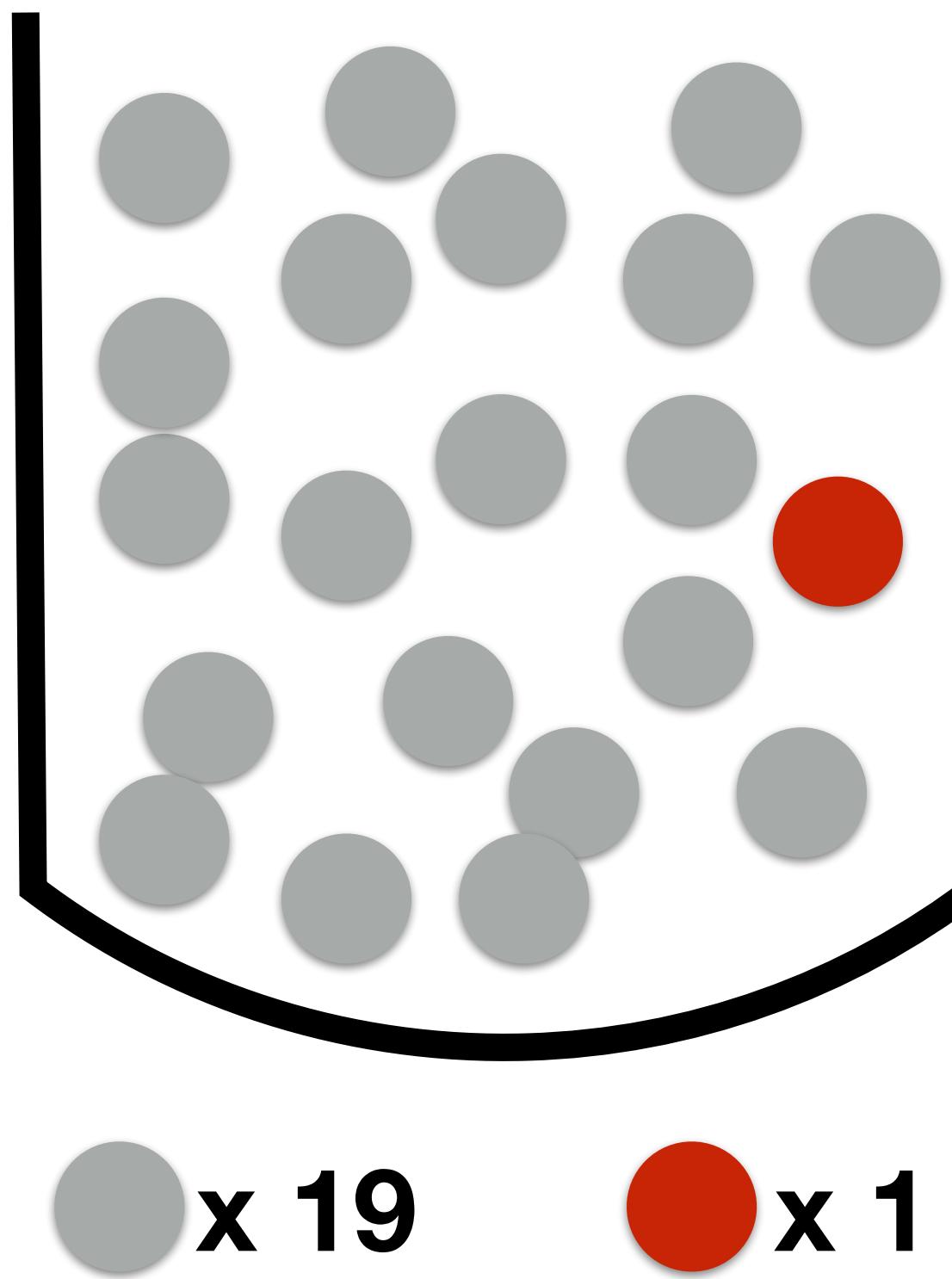
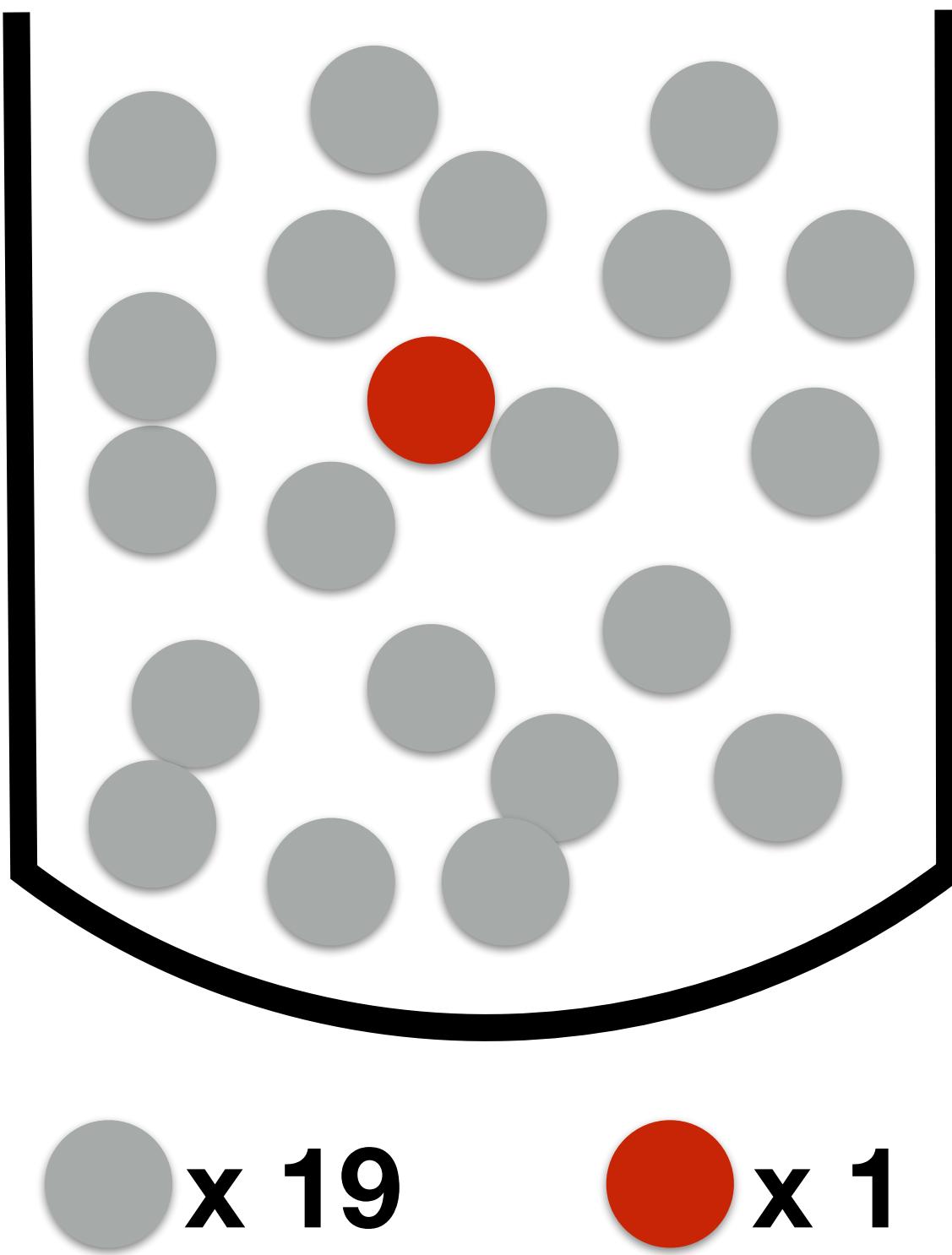
● x 19

● x 1



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

$$\begin{aligned}1 - (1 - 0.05)^2 \\= 0.0975\end{aligned}$$



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

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

Psychological Science
22(11) 1359–1366
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
<http://pss.sagepub.com>


Joseph P. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

¹The Wharton School, University of Pennsylvania, and ²Haas School of Business, University of California, Berkeley

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

labphon 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

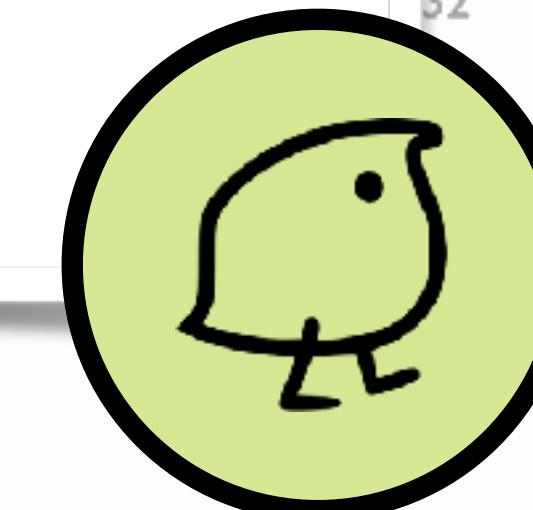
False-Positive Flexibility Allows P-Hacking

Researcher degrees of freedom in phonetic research

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

can be a problem,
and the research

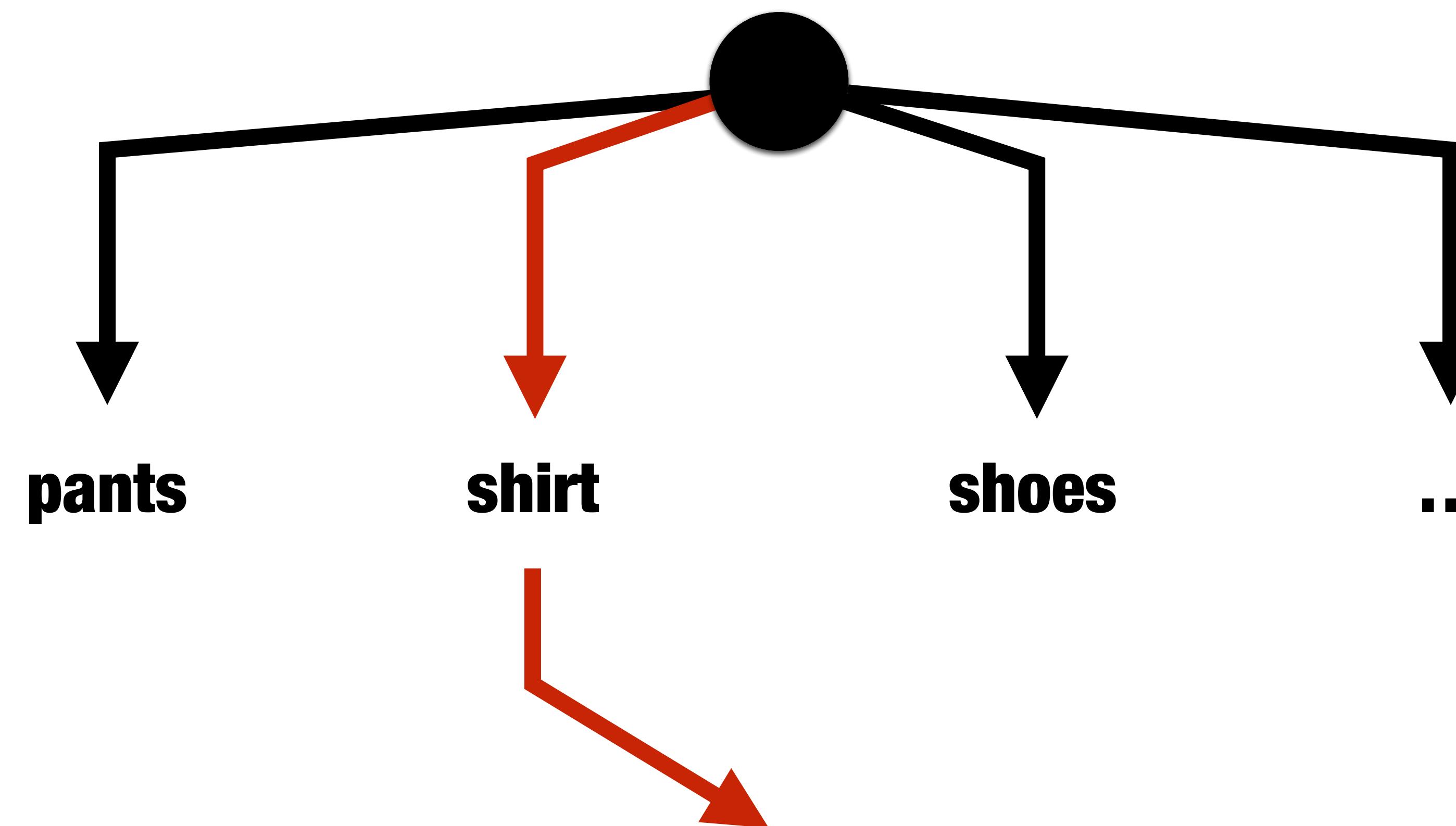
ns.nav
32



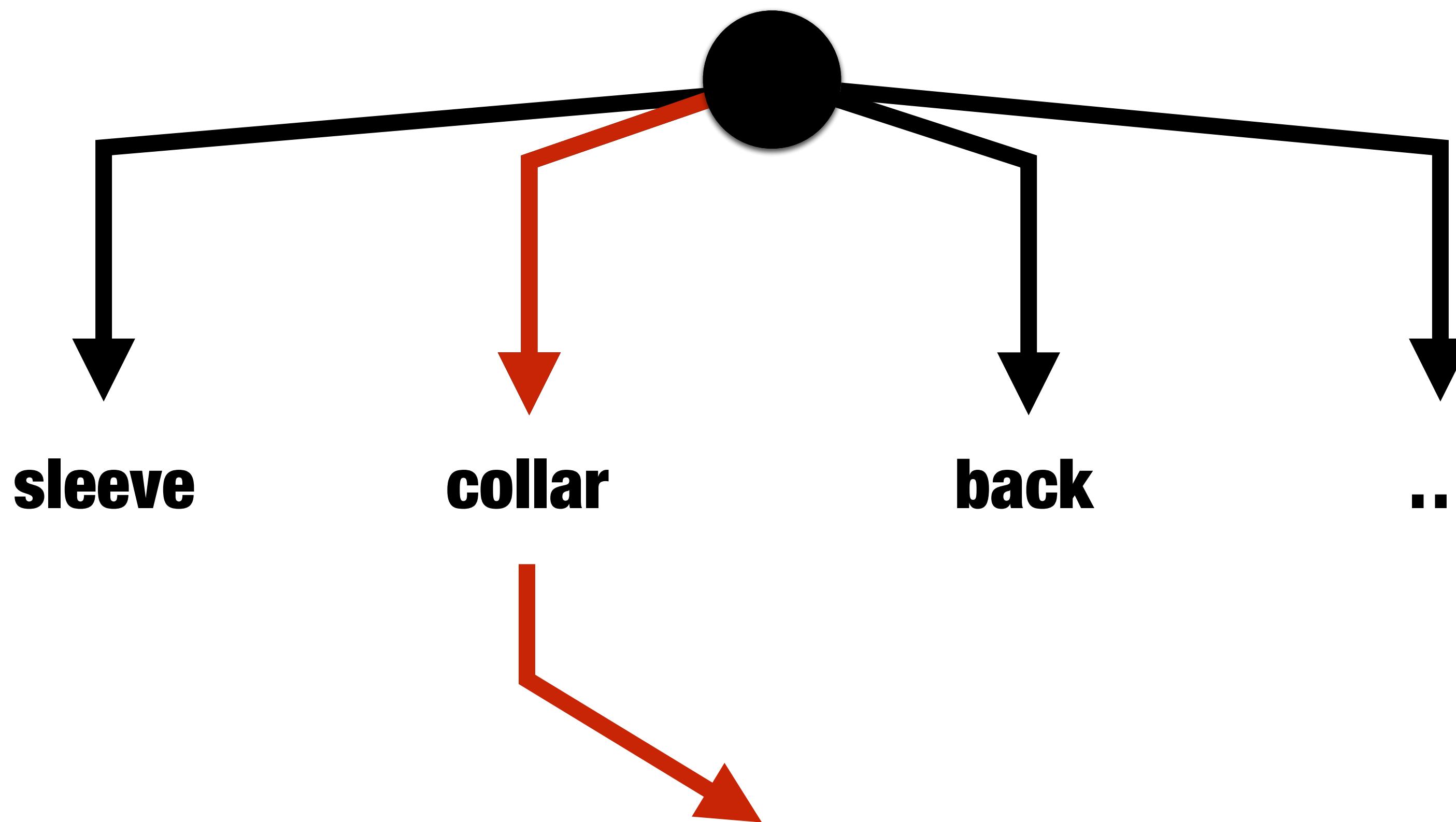
Joseph P. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

¹The Wharton School, University of Pennsylvania, and ²Haas School of Business, University of California, Berkeley

choose a piece of clothing

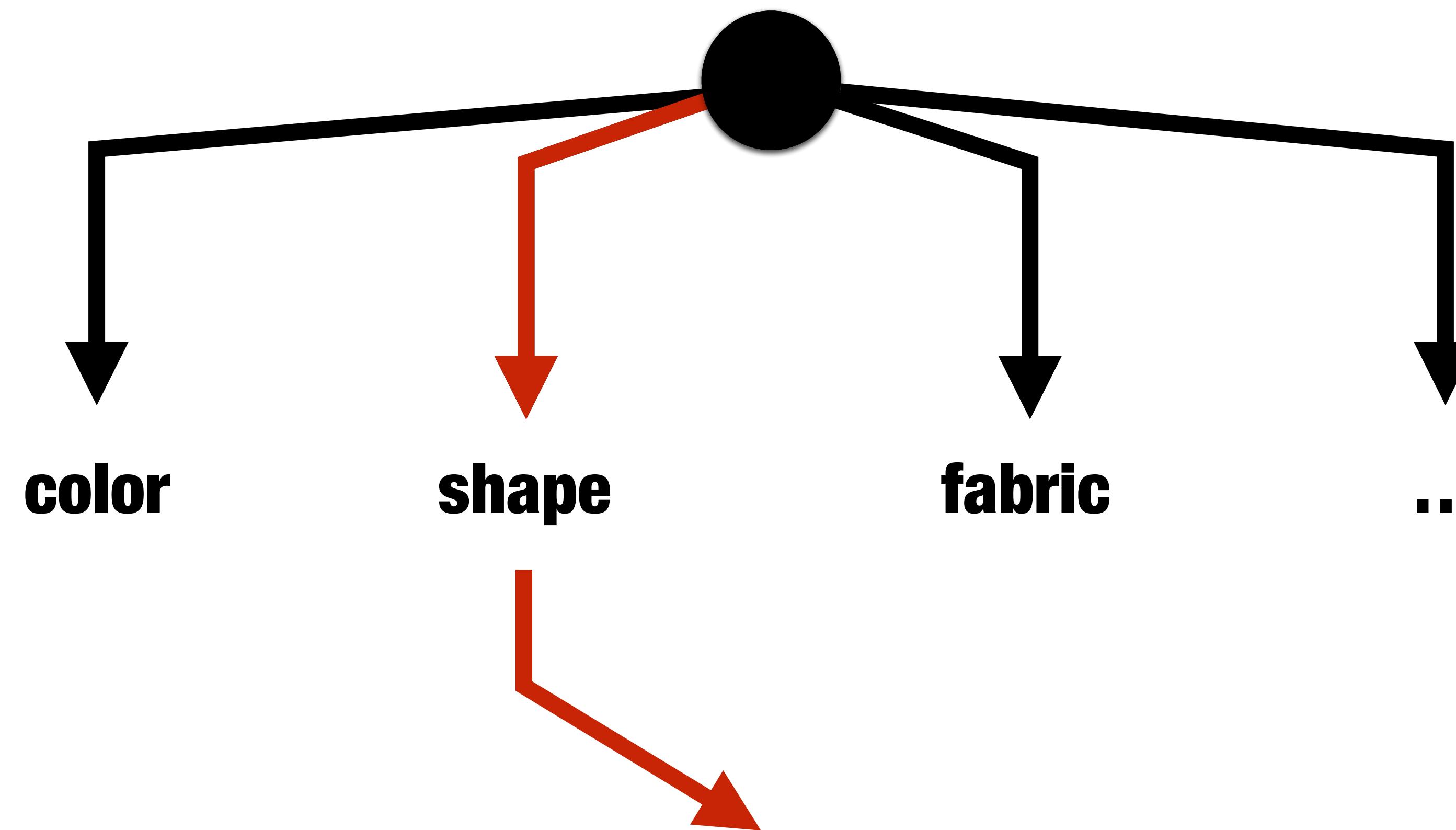


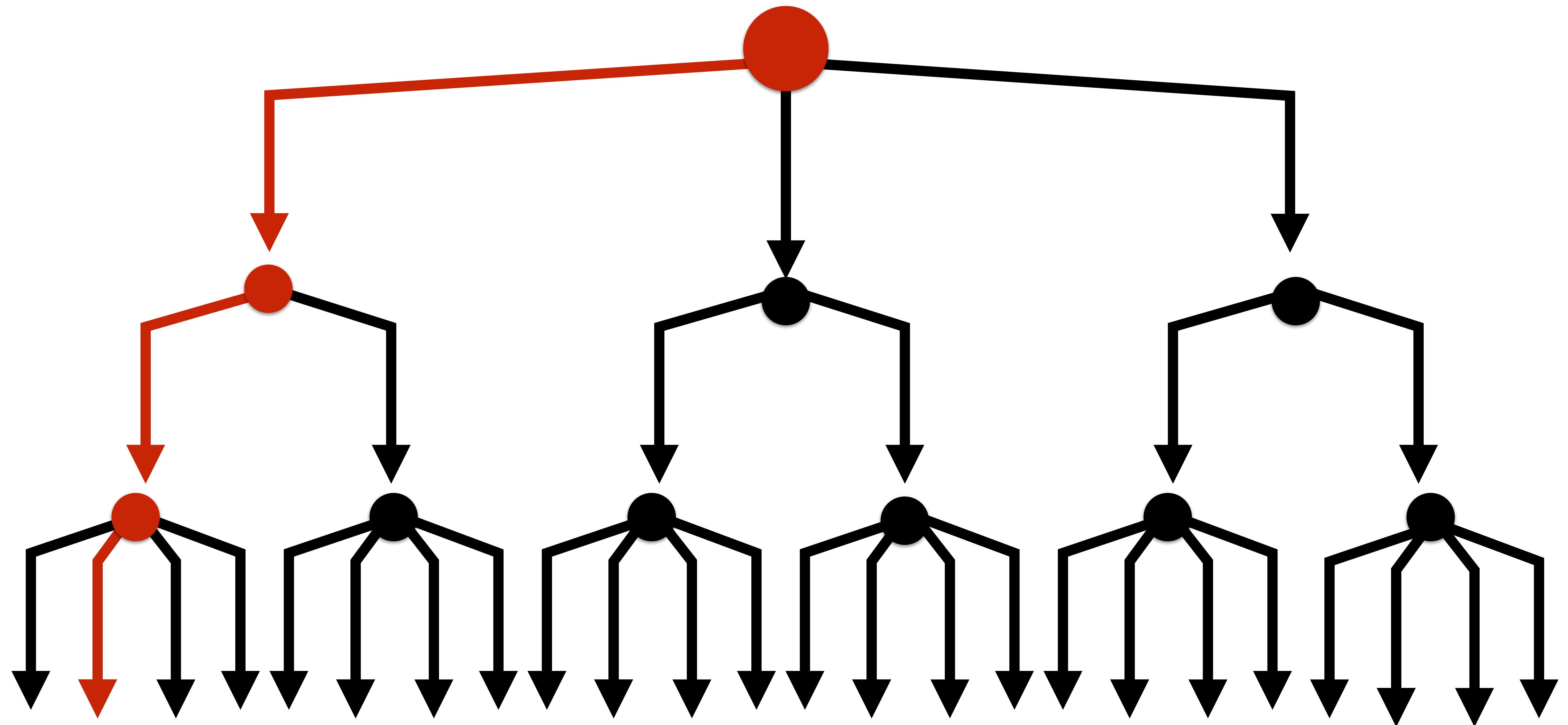
choose part of clothing

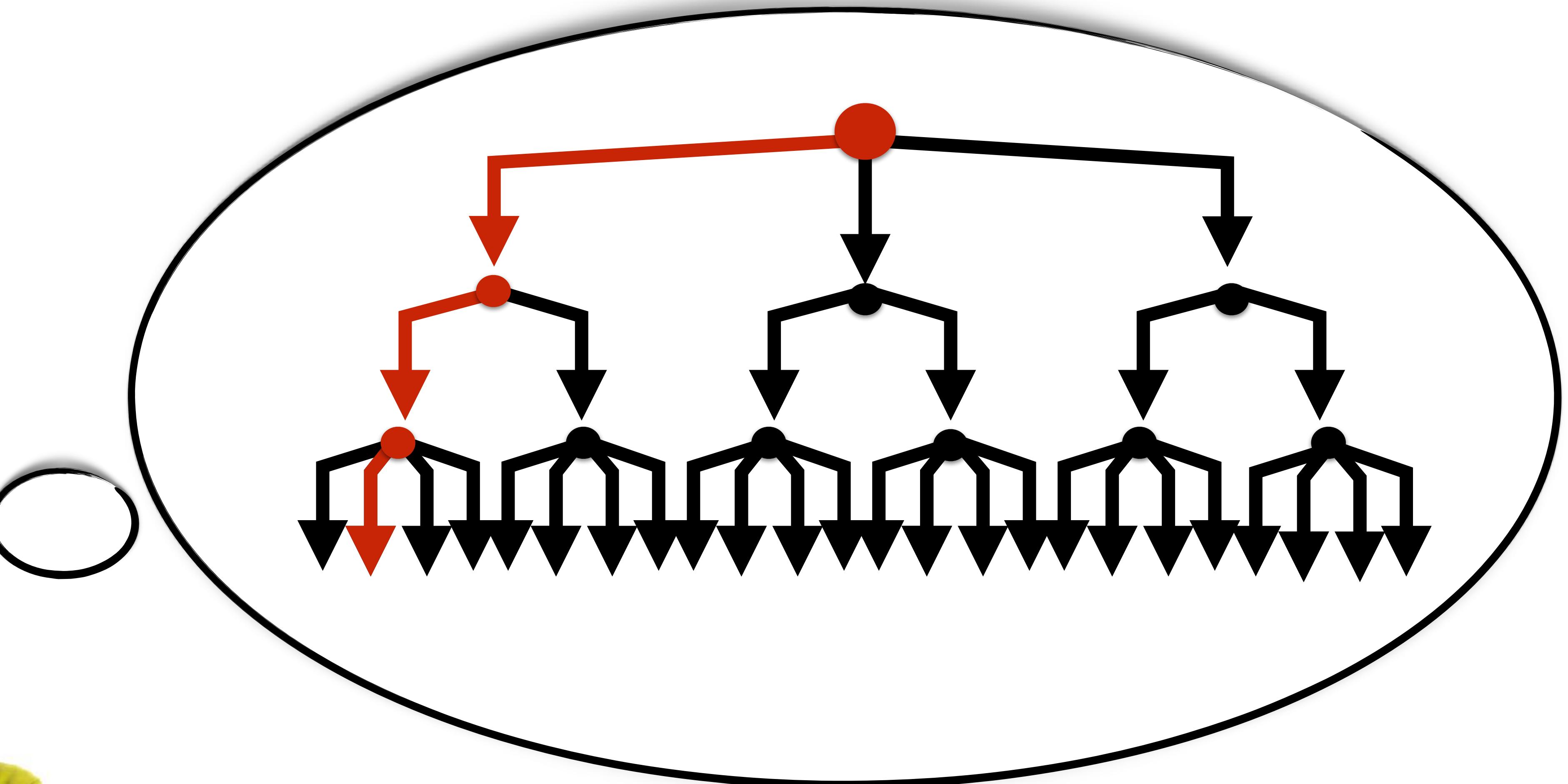


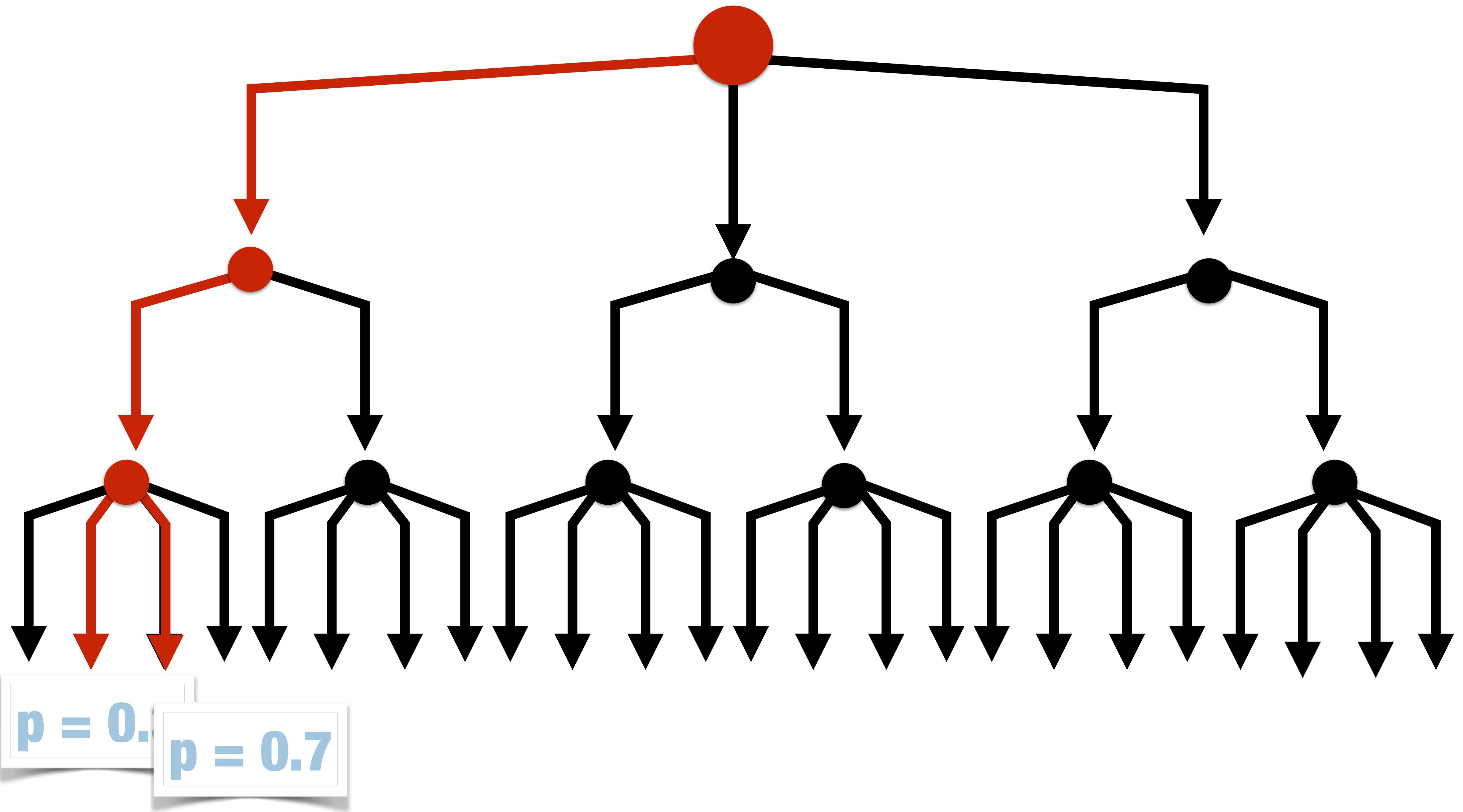


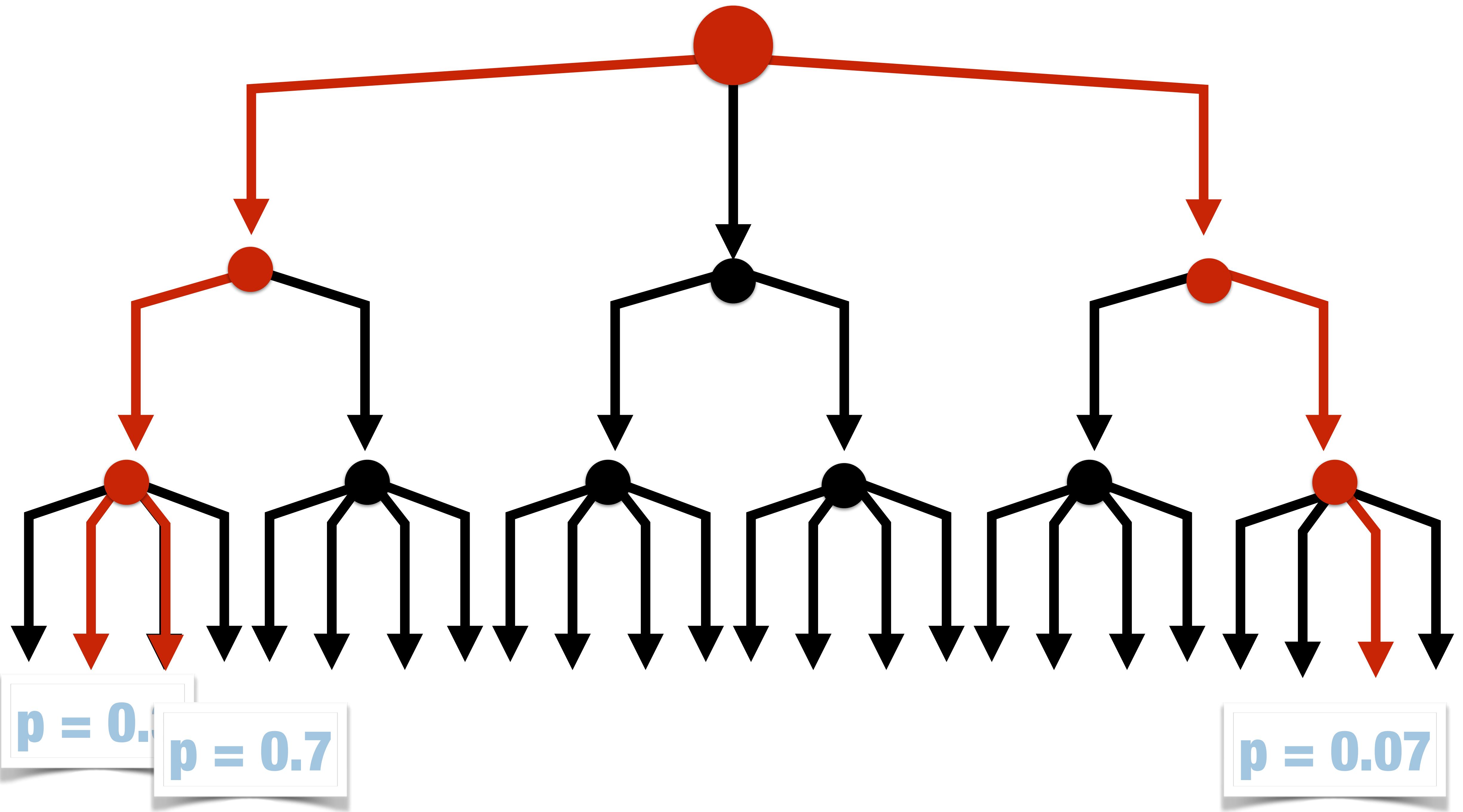
choose concrete operationalization

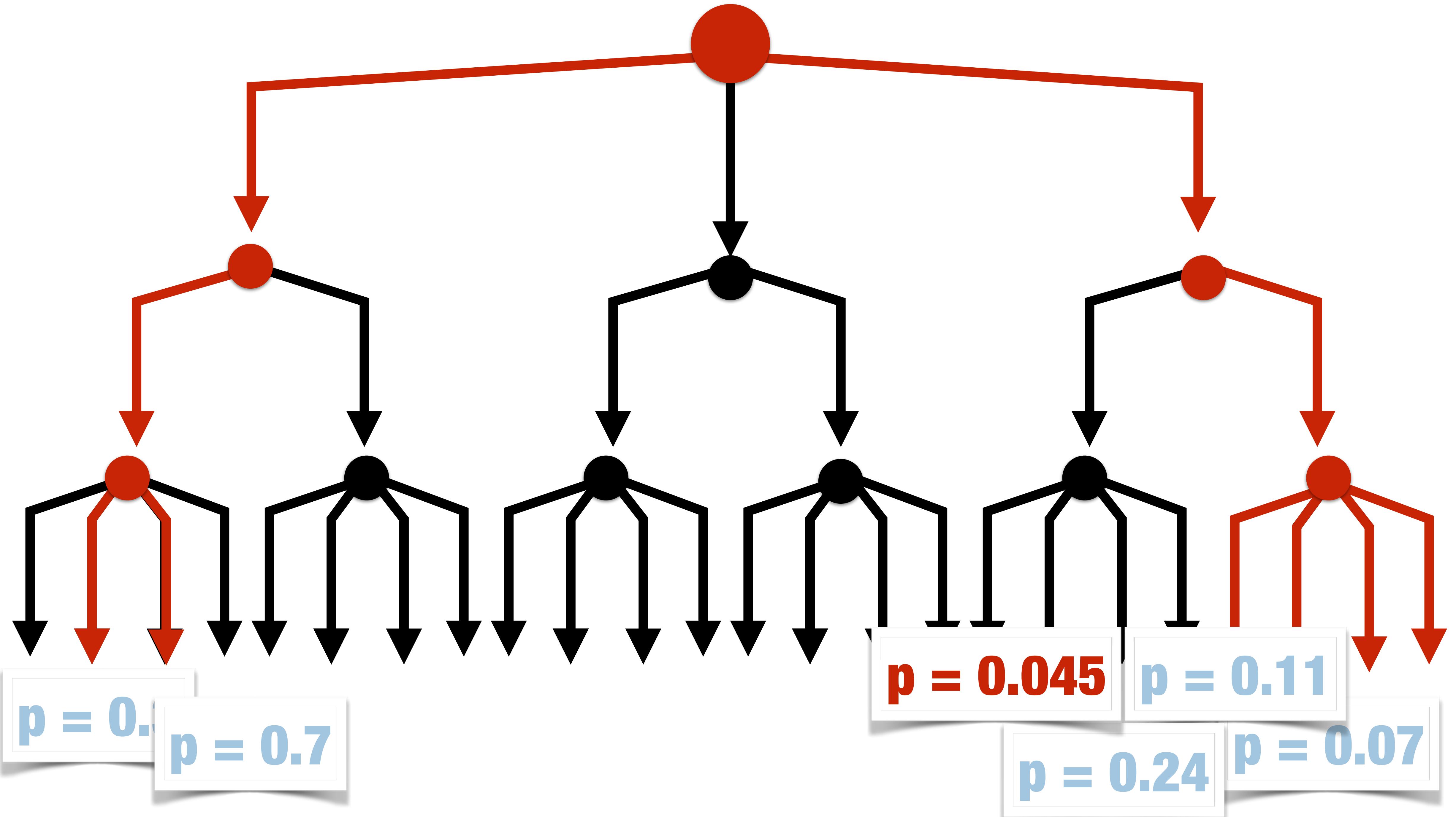


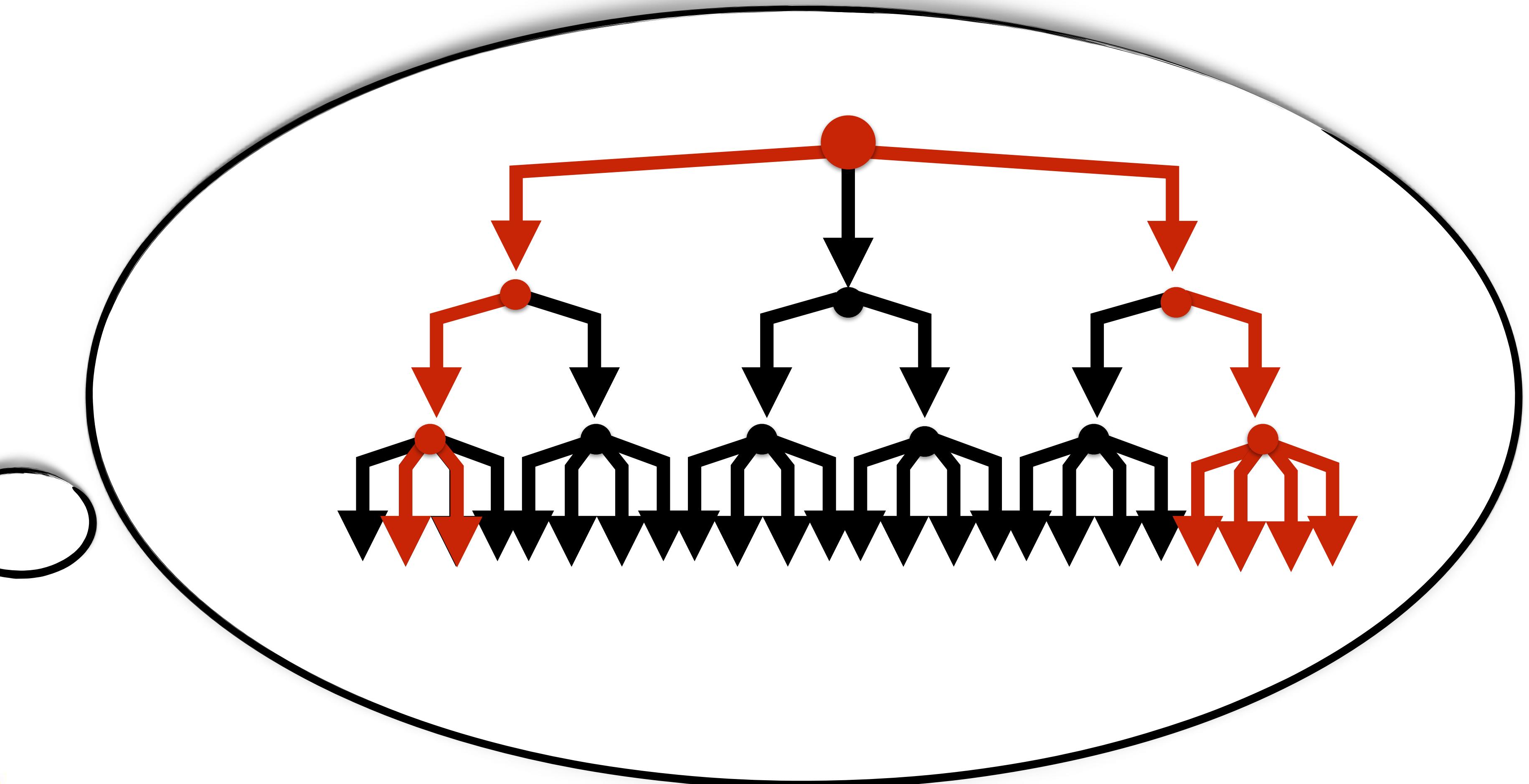


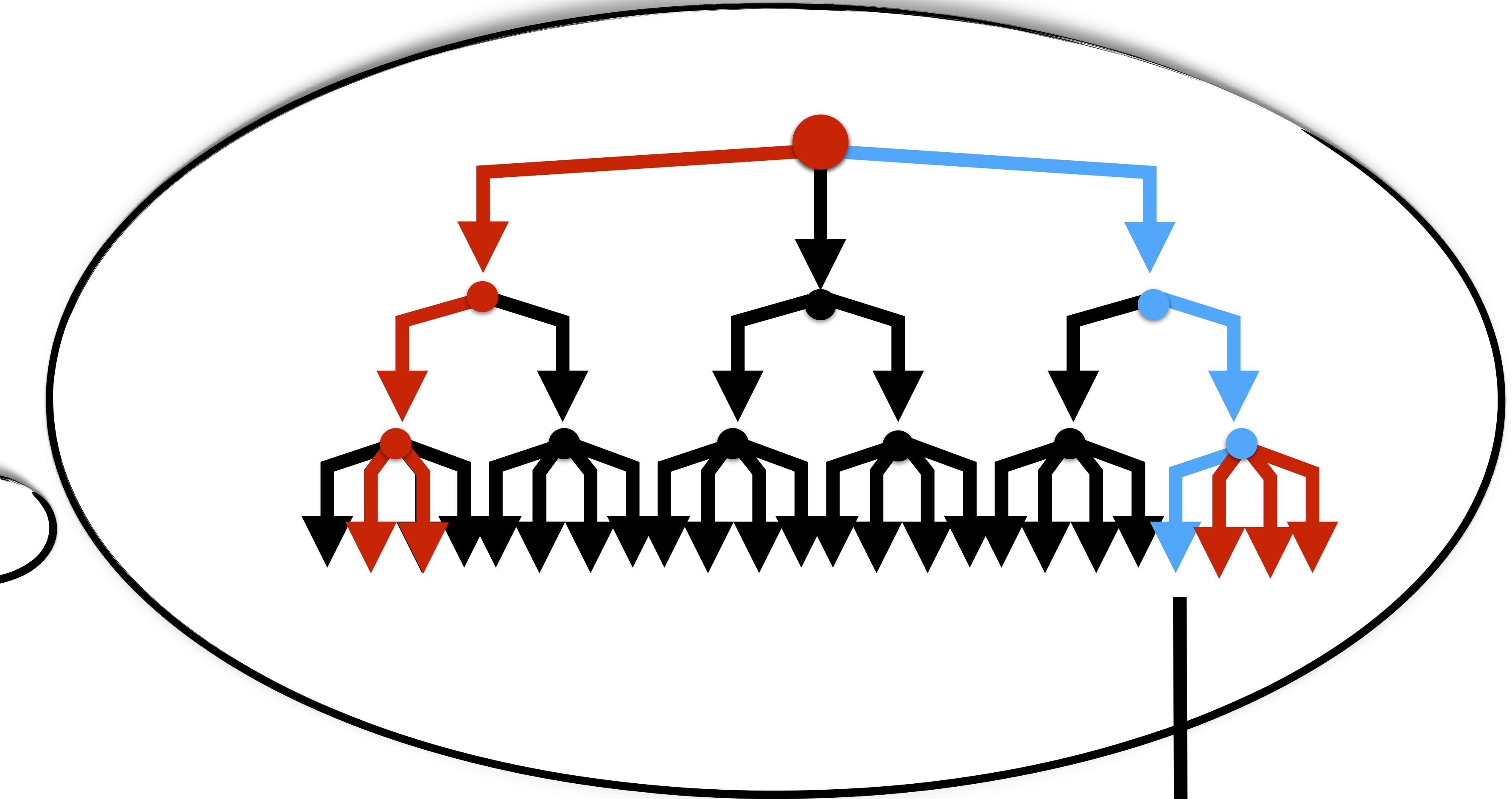






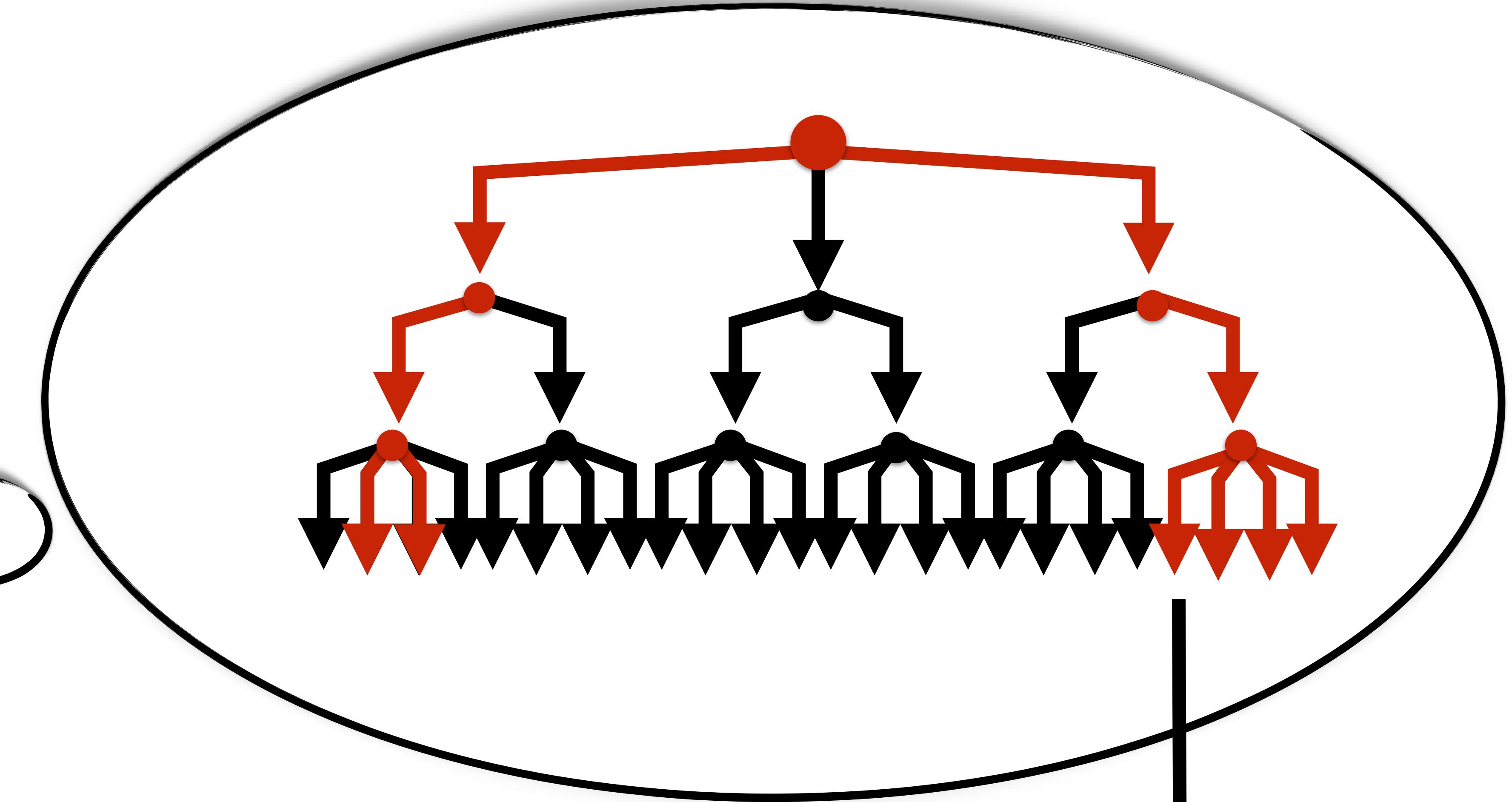






**Report only
successful
pipeline**

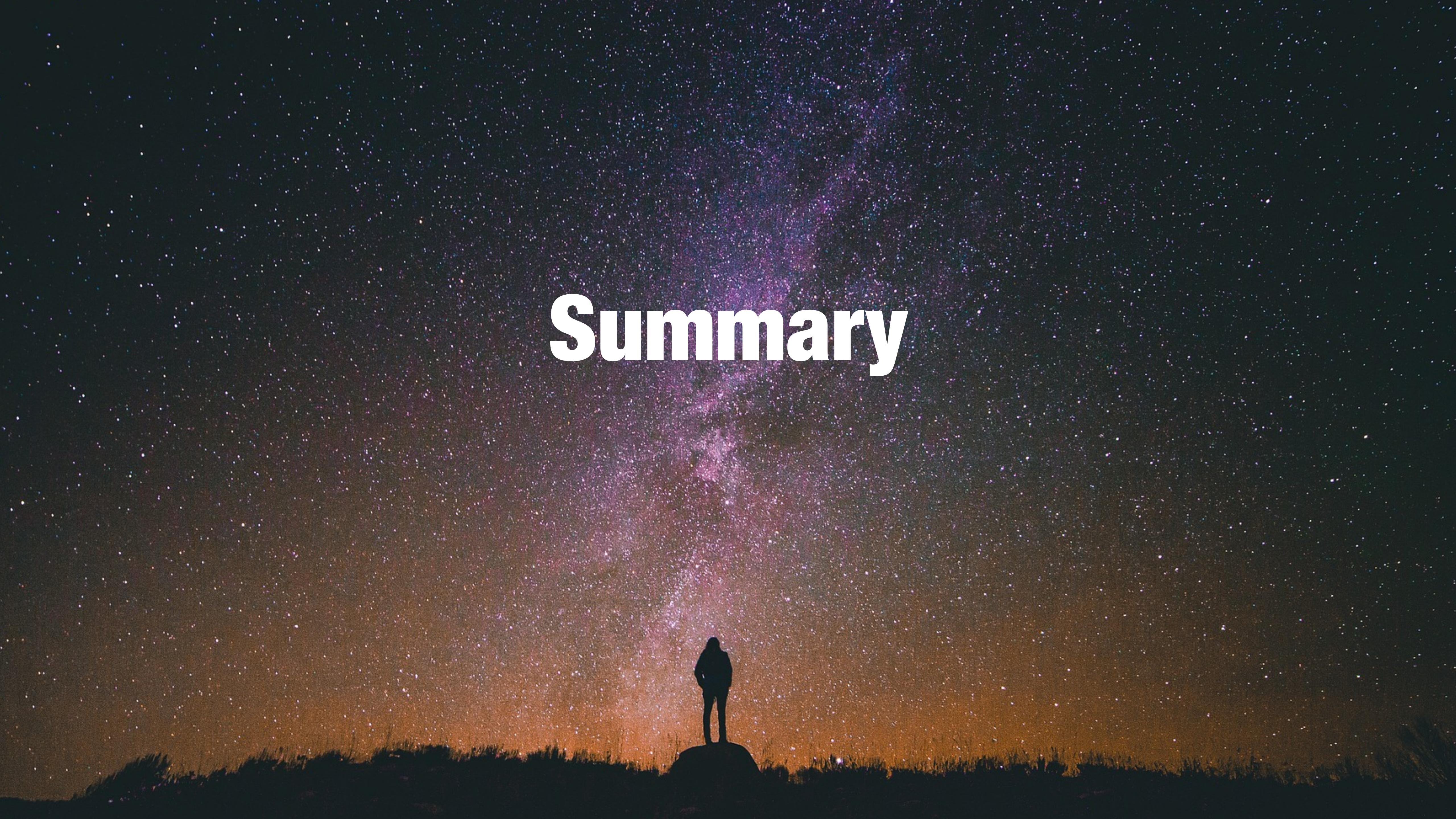




**Report all
pipelines but
do not correct**

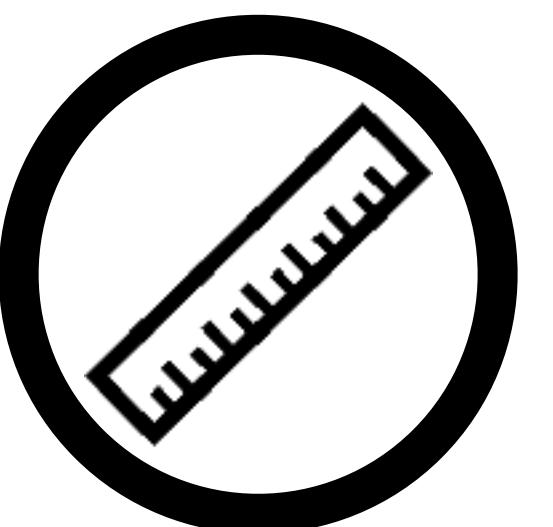


Summary





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

dichotomous decision making

is subject to false positives and
false negatives

analytical flexibility

can amplify human error and bias

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.
- Vasisht, 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.