

Statistical inference and things that can go wrong

Timo B. Roettger



@TimoRoettger



**election
polls**



**human
cognition**



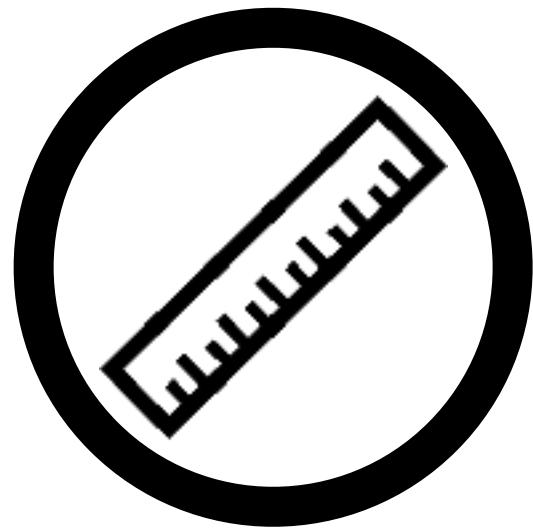
**scientific
publications**



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



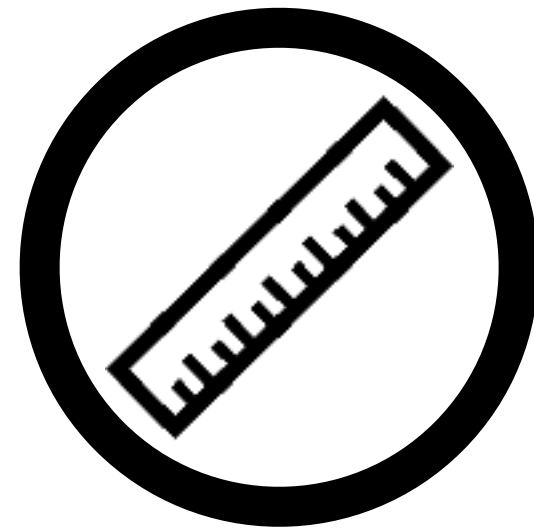
Is there a relationship
between **autism** and
vaccination?



measure some stuff...



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



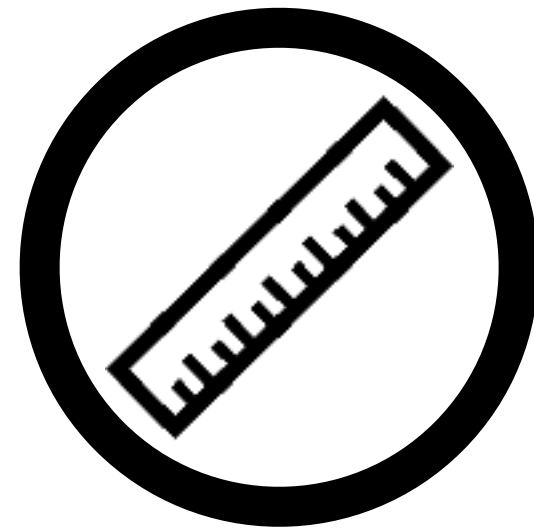
measure some stuff...



Evaluate that
measured stuff
using **statistical
inference**



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



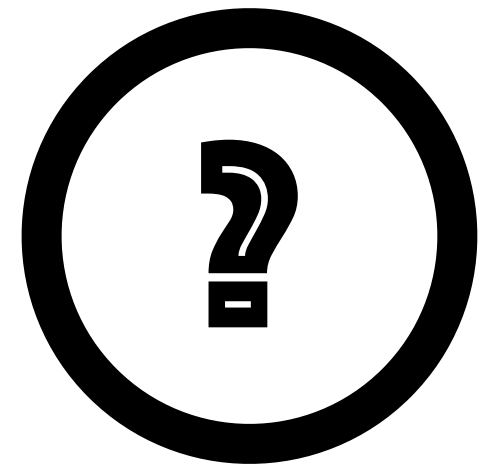
measure some stuff...



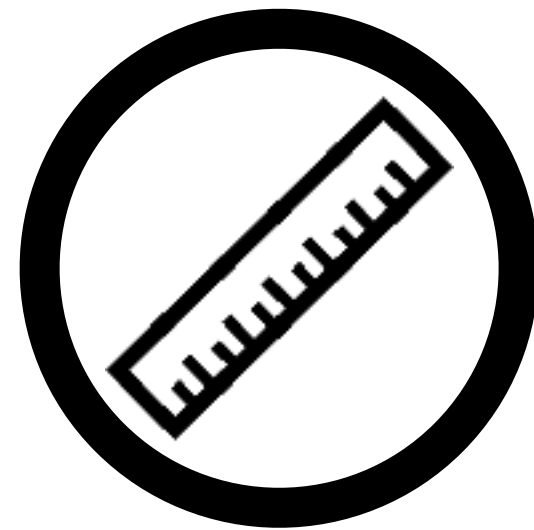
Evaluate that
measured stuff
using **statistical
inference**

Publish findings in
a scientific journal





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

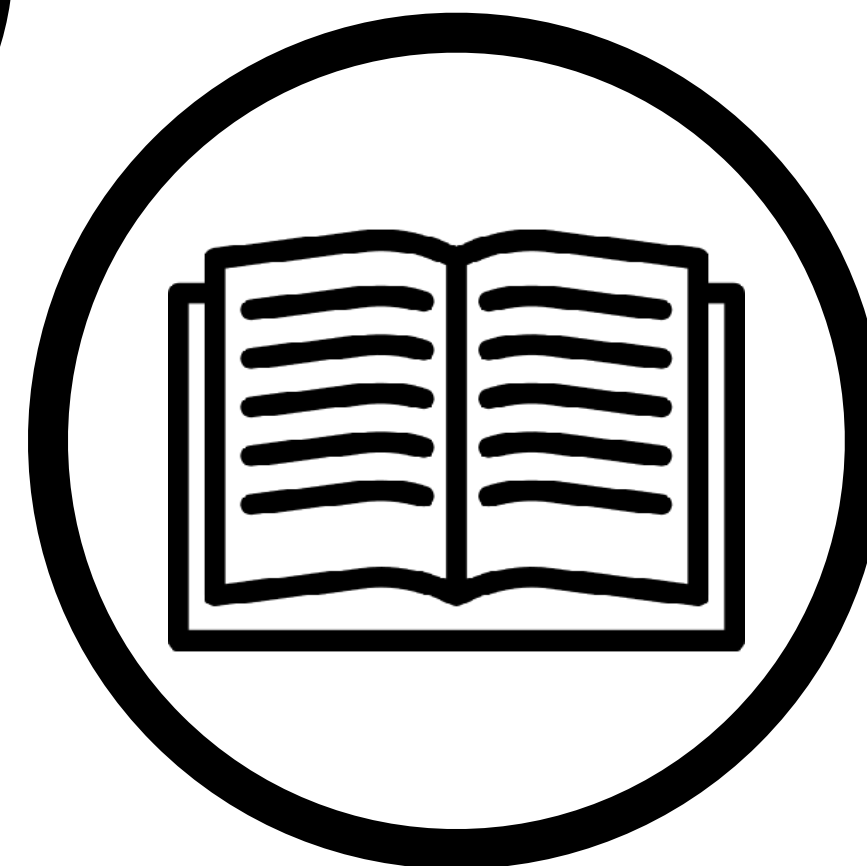


measure some stuff...



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
7,423

LIKES
15,172



2:09 PM - 13 Jan 2020



912



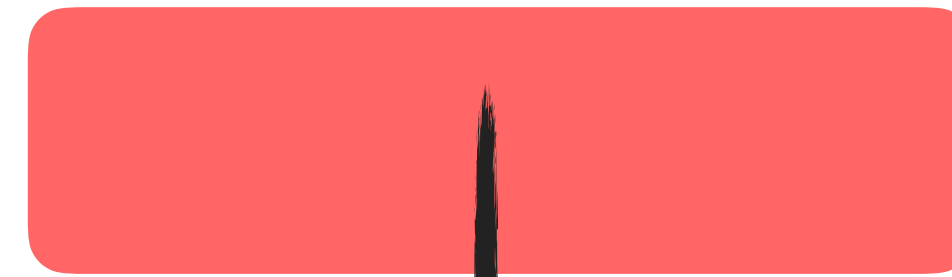
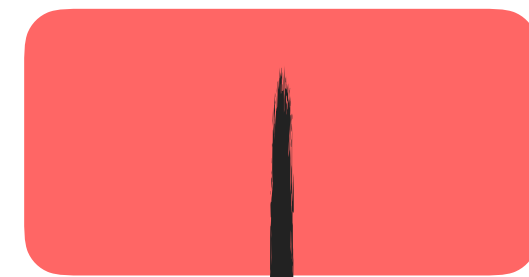
7K



15K



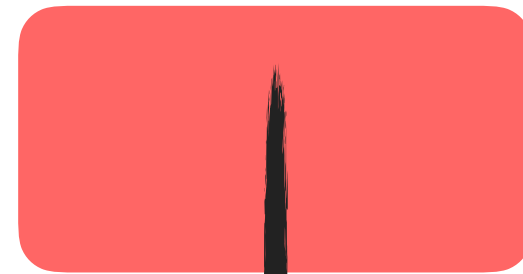
Statistical inference and NHST



the **infamous**
p-value

Things that
can go
terribly wrong

Statistical inference and NHST



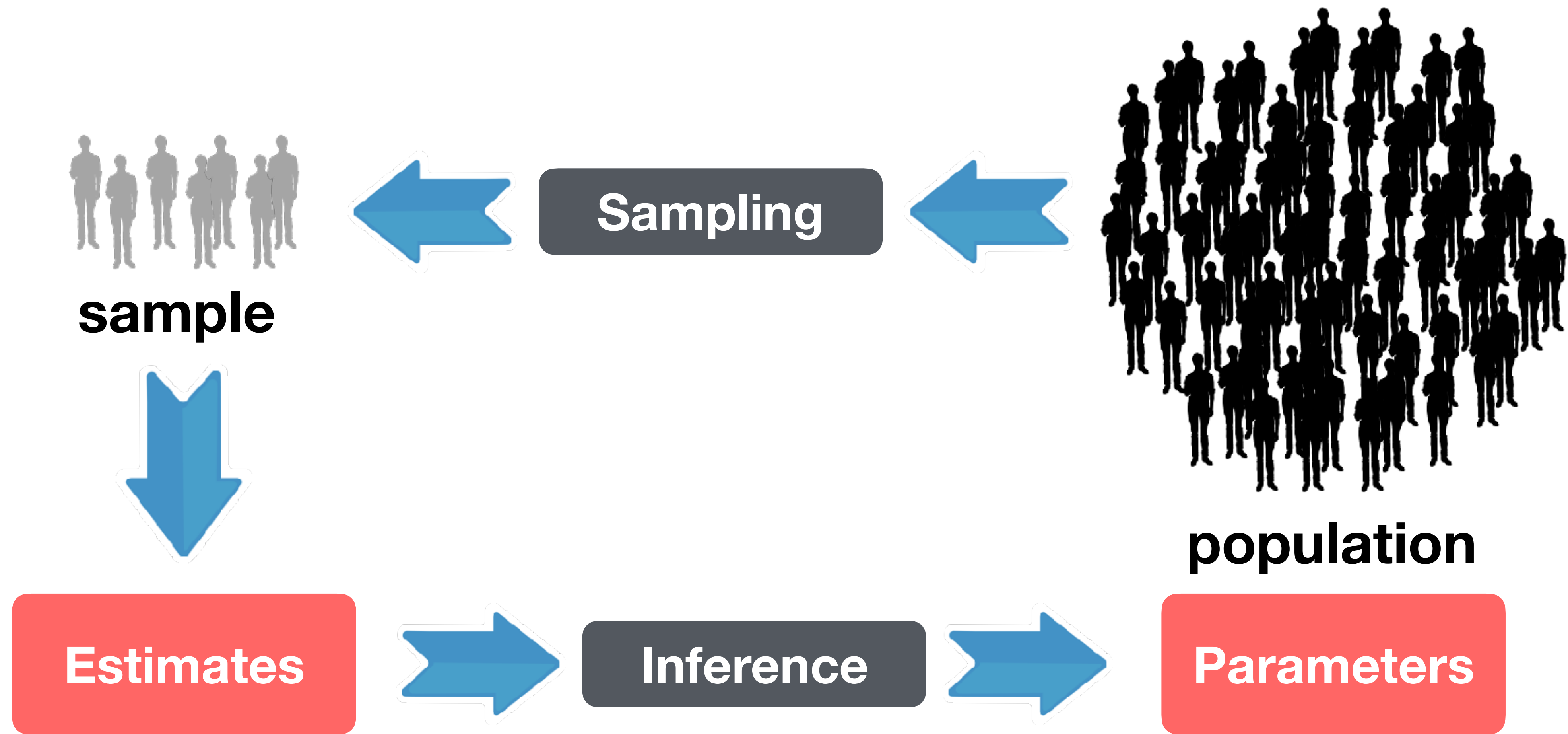
the **infamous**
p-value

Things that
can go
terribly wrong

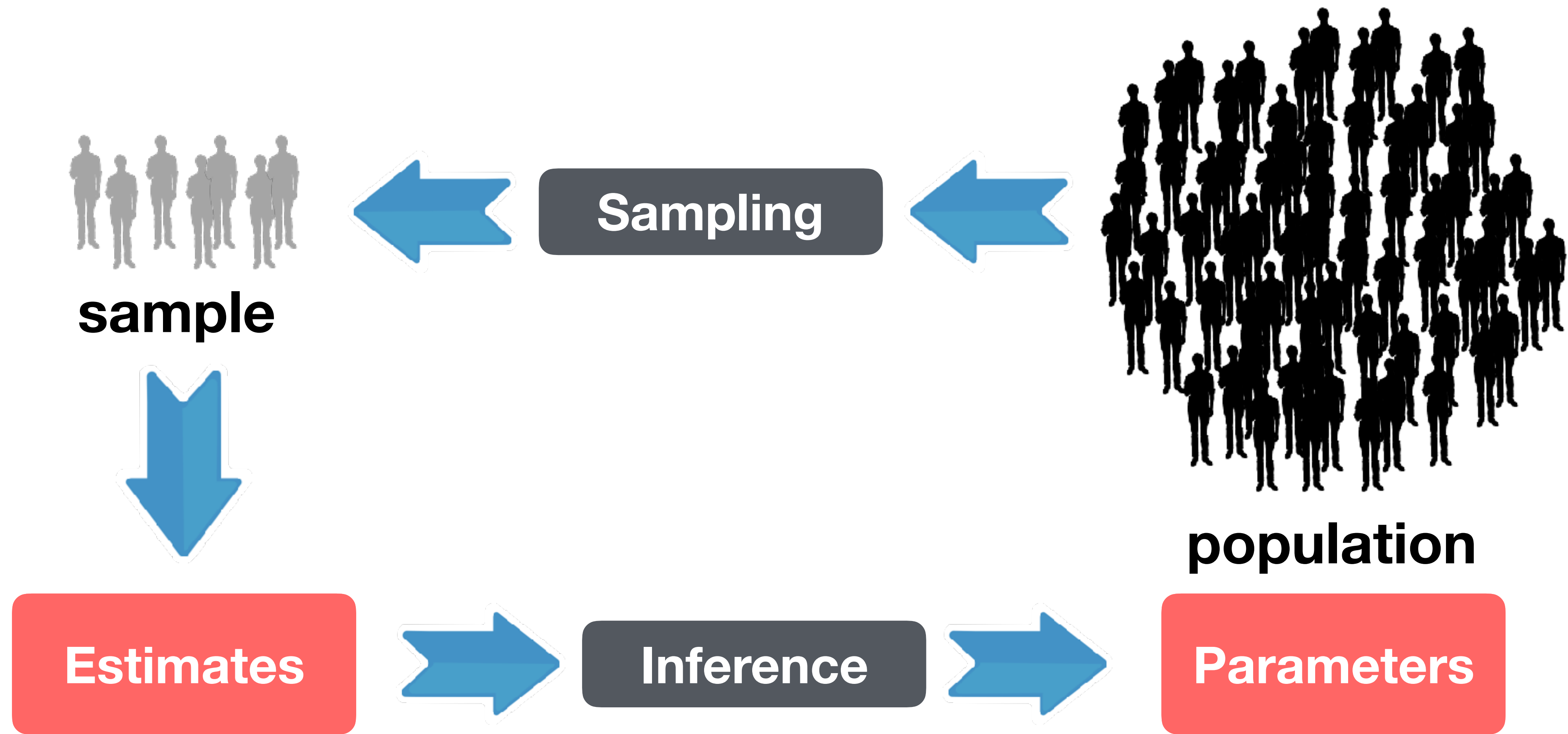
Statistical inference



Statistical inference



Statistical inference



NHST

Null-Hypothesis Significance Testing

NHST

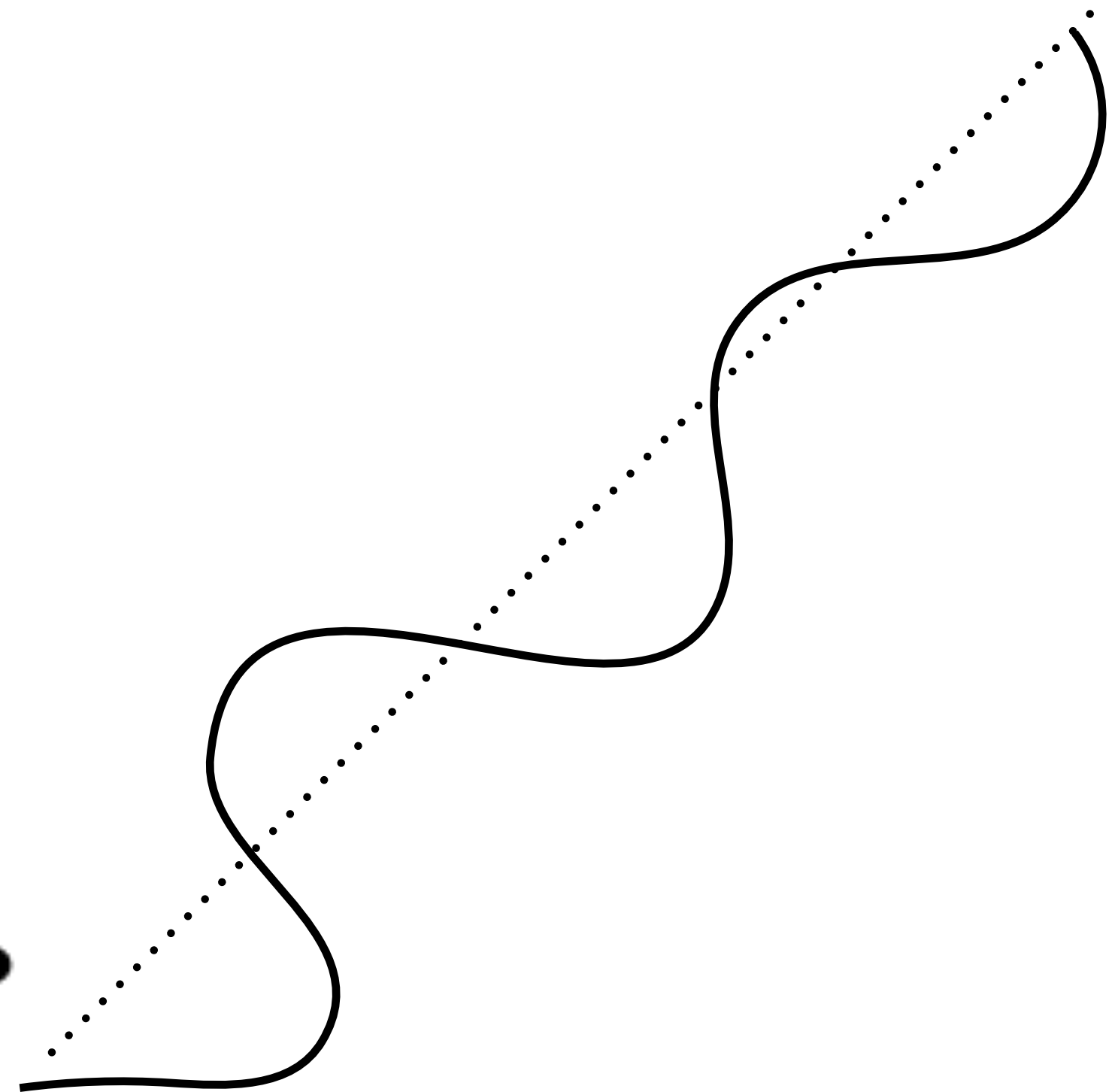
Null-Hypothesis Significance Testing

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

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.

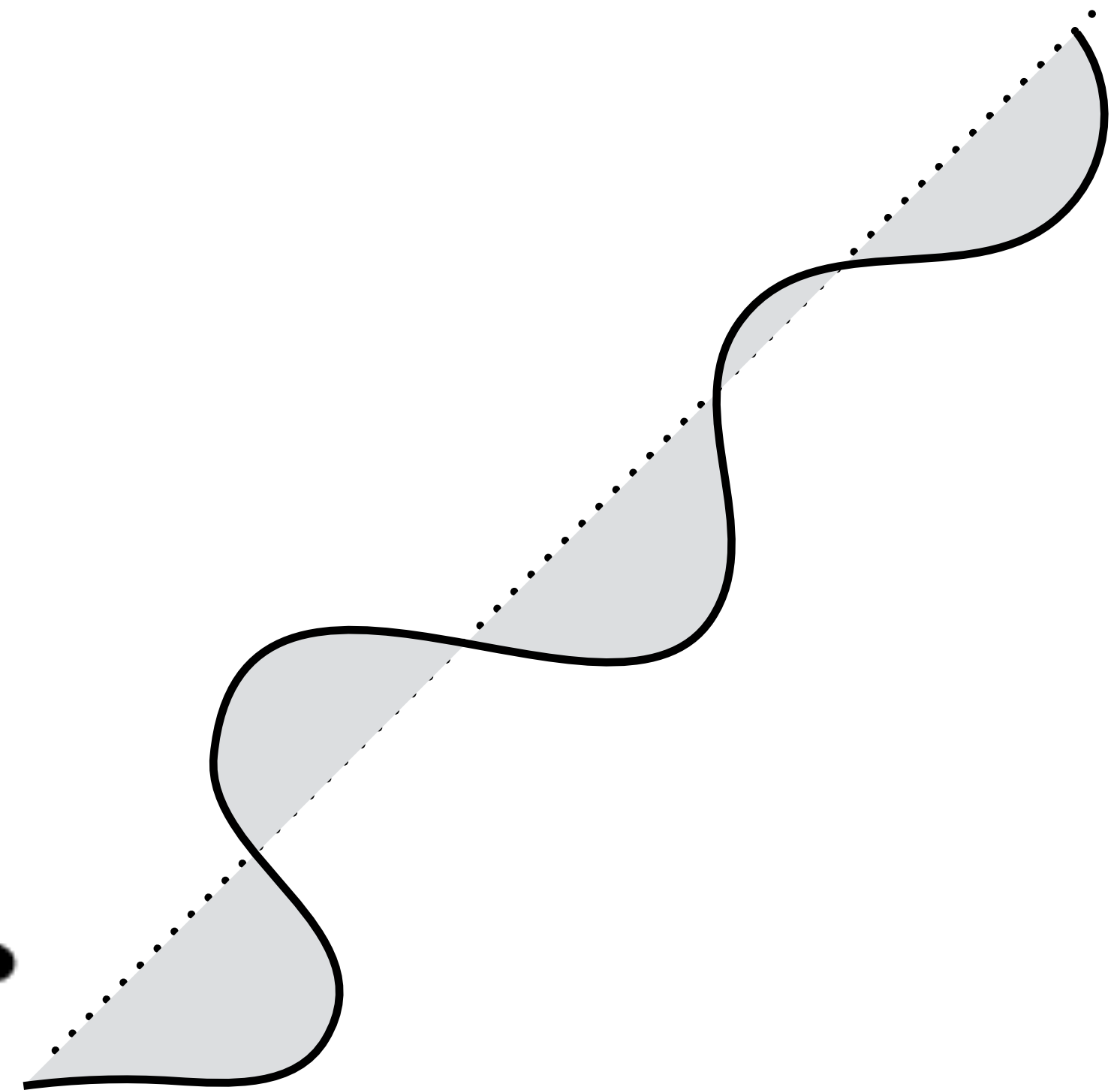


H_a : Drunken people diverge more from a straight line than sober people.

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

the larger t ,
the smaller p

difference between groups

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

standard error (SE)

“noise”

sample size

Measuring the compatibility of the data

the larger t ,
the smaller p

difference between groups

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

standard error (SE)

“noise”

sample size

The diagram illustrates the t-test formula with several annotations. The variable t is highlighted in a light blue box. The numerator, $\bar{x}_1 - \bar{x}_2$, is enclosed in a light red box and labeled 'difference between groups' in red text above it. The denominator is a square root expression. Inside the square root, the standard deviation SD is in a light green box and labeled '“noise”' in green text to its right. Below SD , the sample size N is also in a light green box and labeled 'sample size' in green text to its right. A large green curly bracket on the left side of the denominator is labeled 'standard error (SE)' in green text.

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

The diagram illustrates the calculation of a t-value. On the left, a blue box contains the value '9.8', with the label 't value' above it. This is followed by an equals sign. To the right of the equals sign is a fraction. The numerator of the fraction is '4 - 2', enclosed in a pink box, with the label 'difference between groups' above it. The denominator is a square root symbol containing a fraction. The numerator of this inner fraction is '1', enclosed in a light green box, with the label '“noise”' to its right. The denominator of the inner fraction is '24', also enclosed in a light green box, with the label 'sample size' to its right. A green curly bracket on the left side of the square root symbol is labeled 'standard error (SE)'.

Measuring the compatibility of the data

difference between groups

t value

2.5

4 - 3.5

“noise”

sample size

standard error (SE)

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

The diagram illustrates the calculation of a t-value. On the left, a blue arrow points down to a light blue box containing the number '2.5', which is labeled 't value' in blue text. To the right of this box is an equals sign. Following the equals sign is a fraction. The numerator of the fraction is a pink box containing '4 - 3.5', with the text 'difference between groups' in red above it. A red arrow points down from this box. The denominator is a green box containing '1' over a horizontal line, with '24' below it. A green bracket to the left of the denominator is labeled 'standard error (SE)' in green text. To the right of the denominator, the text '“noise”' is in green above the '1', and 'sample size' is in green below the '24'. A large black square root symbol is positioned to the left of the denominator's fraction.

Measuring the compatibility of the data

difference between groups

t value

5.5

4 - 3.5

standard error (SE)

0.2

24

"noise"

sample size

The diagram illustrates the calculation of a t-value. The formula is presented as $t = \frac{\text{difference between groups}}{\text{standard error (SE)}}$. The numerator, $4 - 3.5$, is highlighted in a pink box and labeled 'difference between groups' in red text above it. The denominator, $\sqrt{\frac{0.2}{24}}$, is highlighted in a green box and labeled 'standard error (SE)' in green text to its left. The value 0.2 is labeled '"noise"' in green text to its right, and the value 24 is labeled 'sample size' in green text to its right. A blue arrow points upwards to the 't value' label, which is positioned above the calculated t-value of 5.5. A green arrow points downwards from the '"noise"' label.

$$t = \frac{4 - 3.5}{\sqrt{\frac{0.2}{24}}} = 5.5$$

Measuring the compatibility of the data

difference between groups

t value

2.7

4 - 3.5

standard error (SE)

0.2

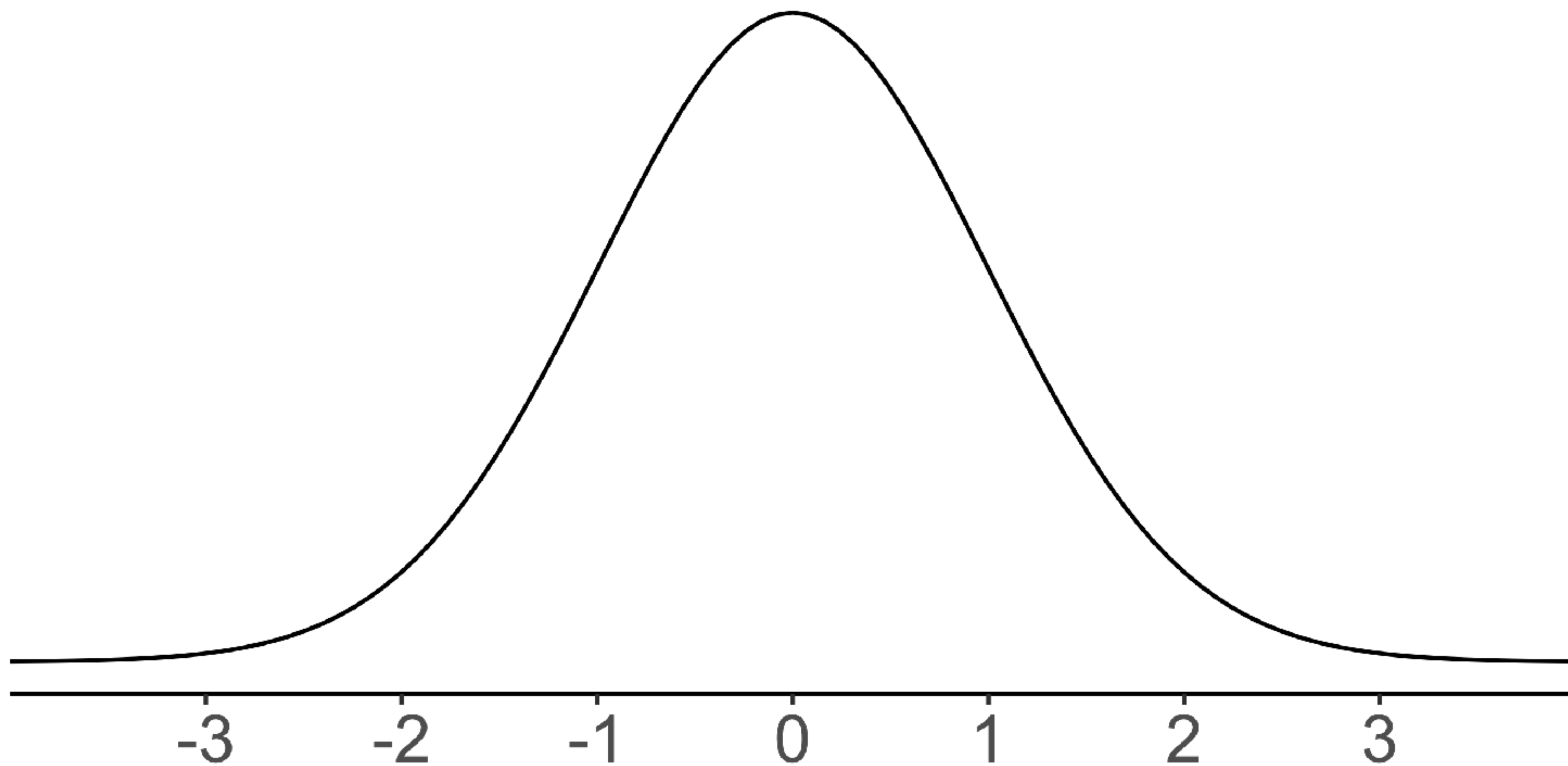
6

“noise”

sample size

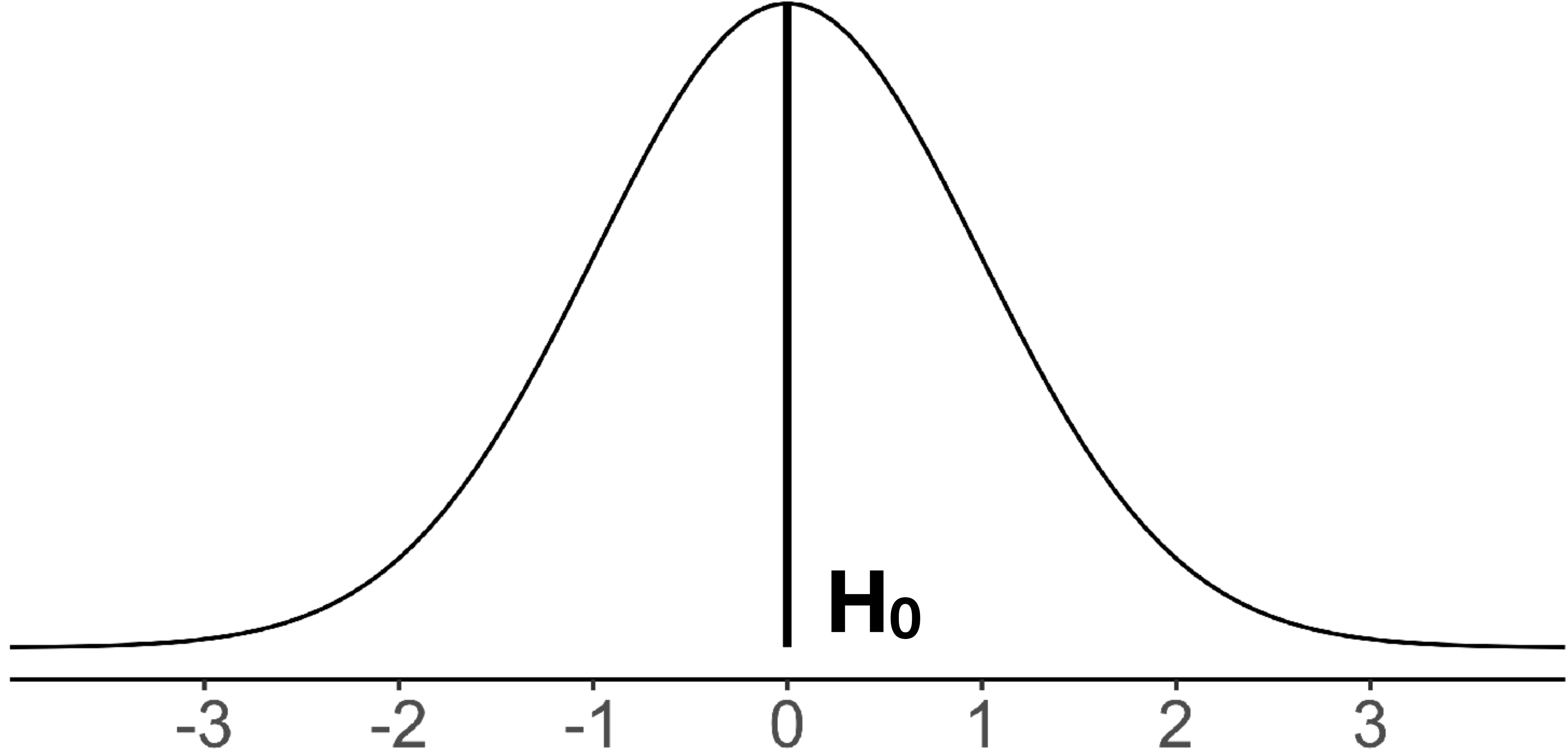
The diagram illustrates the calculation of a t-value. A blue arrow points down to the t-value, which is shown in a light blue box as 2.7. The t-value is equal to the difference between groups (4 - 3.5, shown in a pink box) divided by the standard error (SE). The SE is represented by a green curly bracket and is calculated as the square root of the ratio of 'noise' (0.2, in a light green box) to the sample size (6, in a light green box). A green arrow points down from the sample size.

$$t \text{ value} = \frac{\text{difference between groups}}{\sqrt{\frac{\text{“noise”}}{\text{sample size}}}}$$
$$2.7 = \frac{4 - 3.5}{\sqrt{\frac{0.2}{6}}}$$



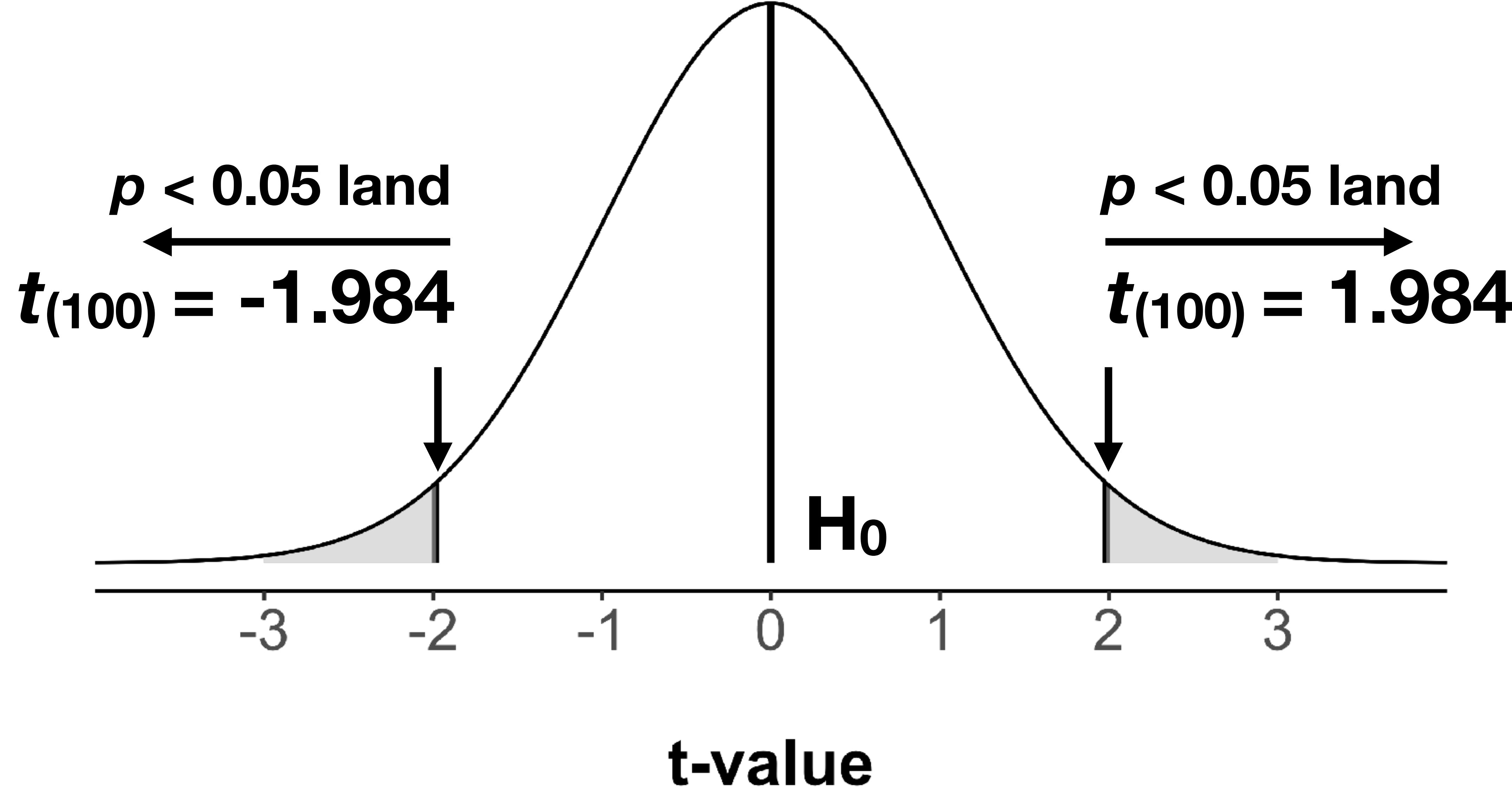
t-value

H_0 : drunk = sober



t-value

H_0 : drunk = sober





drunken



sober

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

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

H_0 : drunk = sober



H_a : drunk > sober*

significant

$p = 0.00524$

$t_{(100)} = 2.6$

H_0

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.

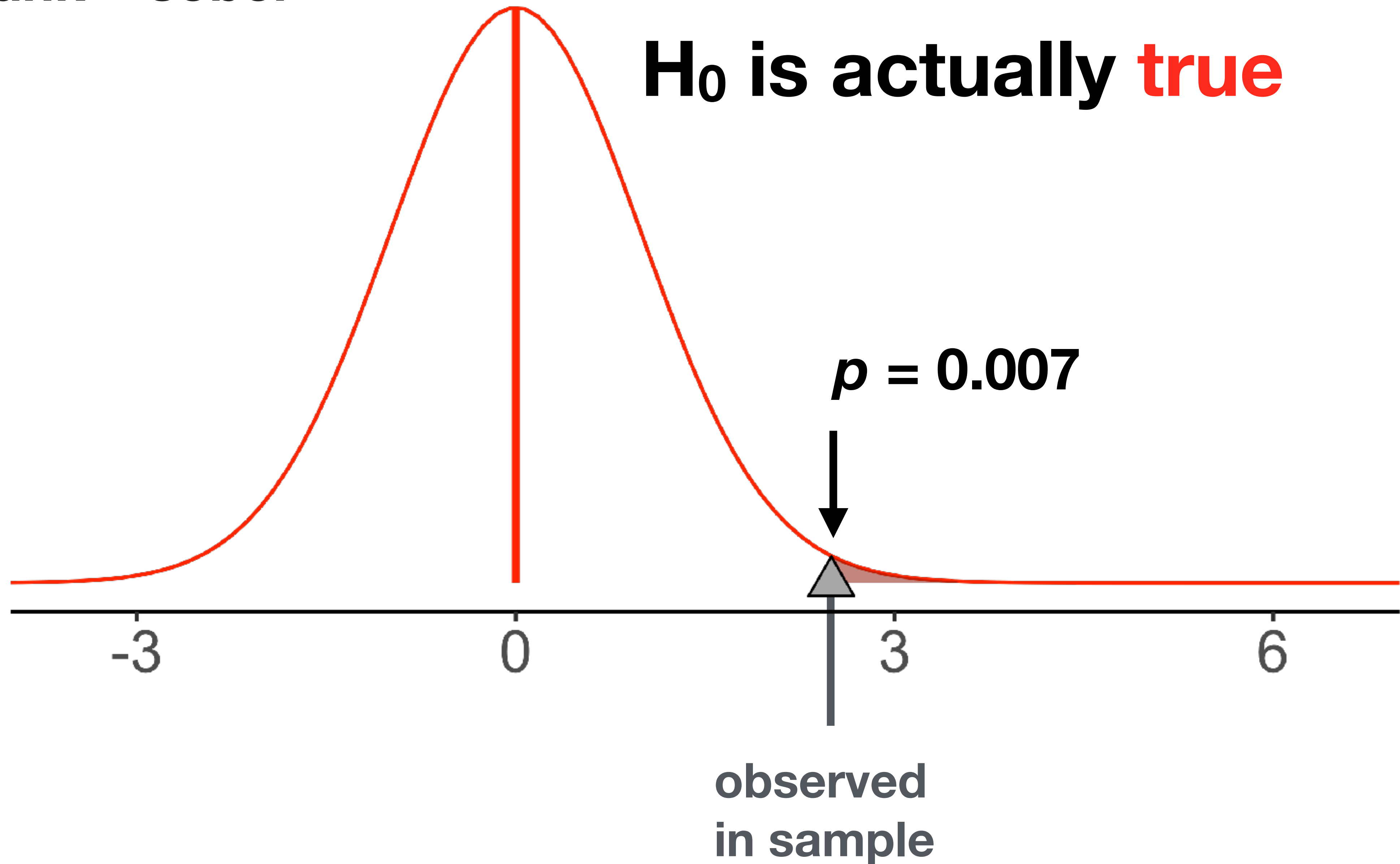
You are **DRUNK**.

False Positive
Type-I error



H_0 : drunk = sober

H_0 is actually **true**

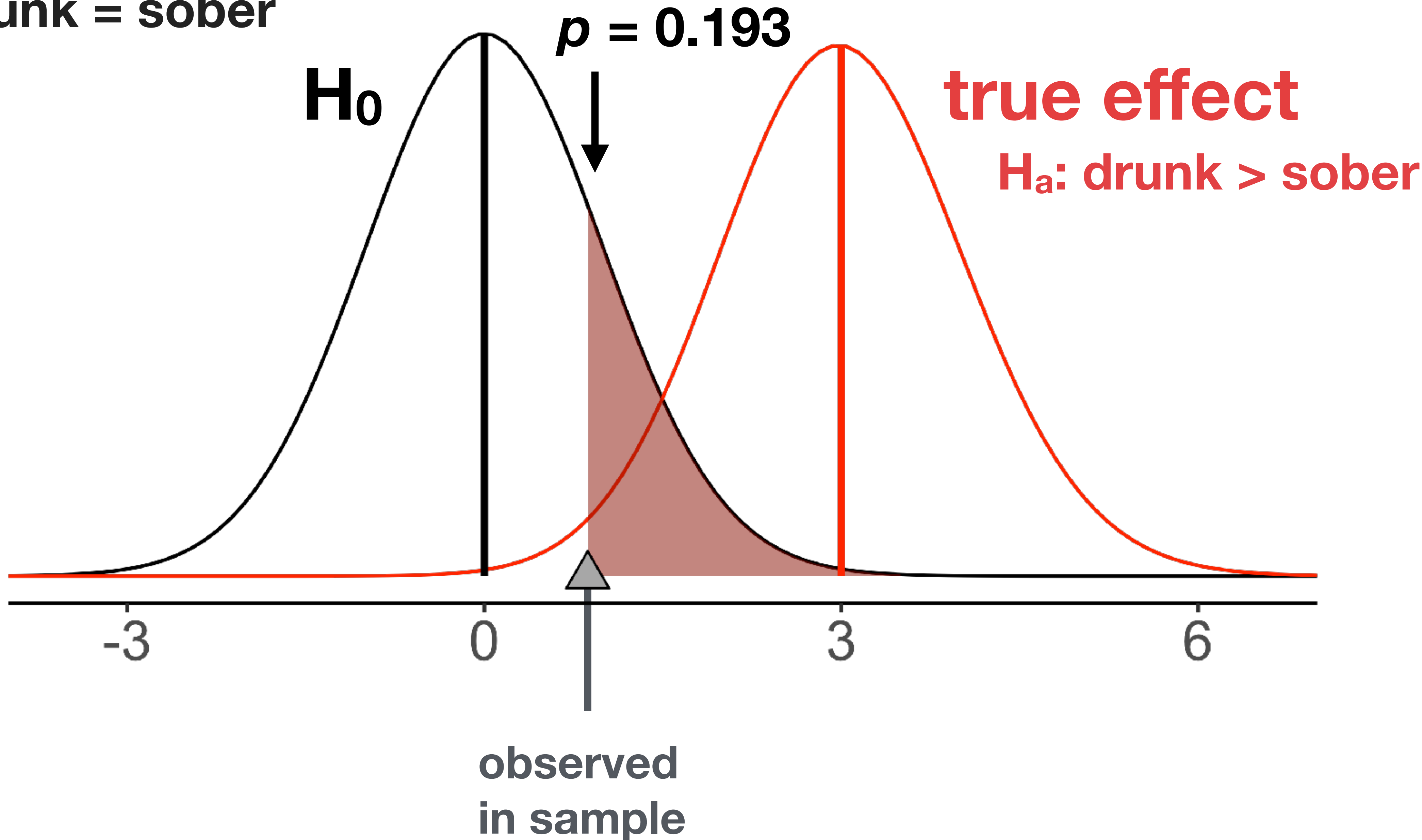


You are **NOT** drunk.

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

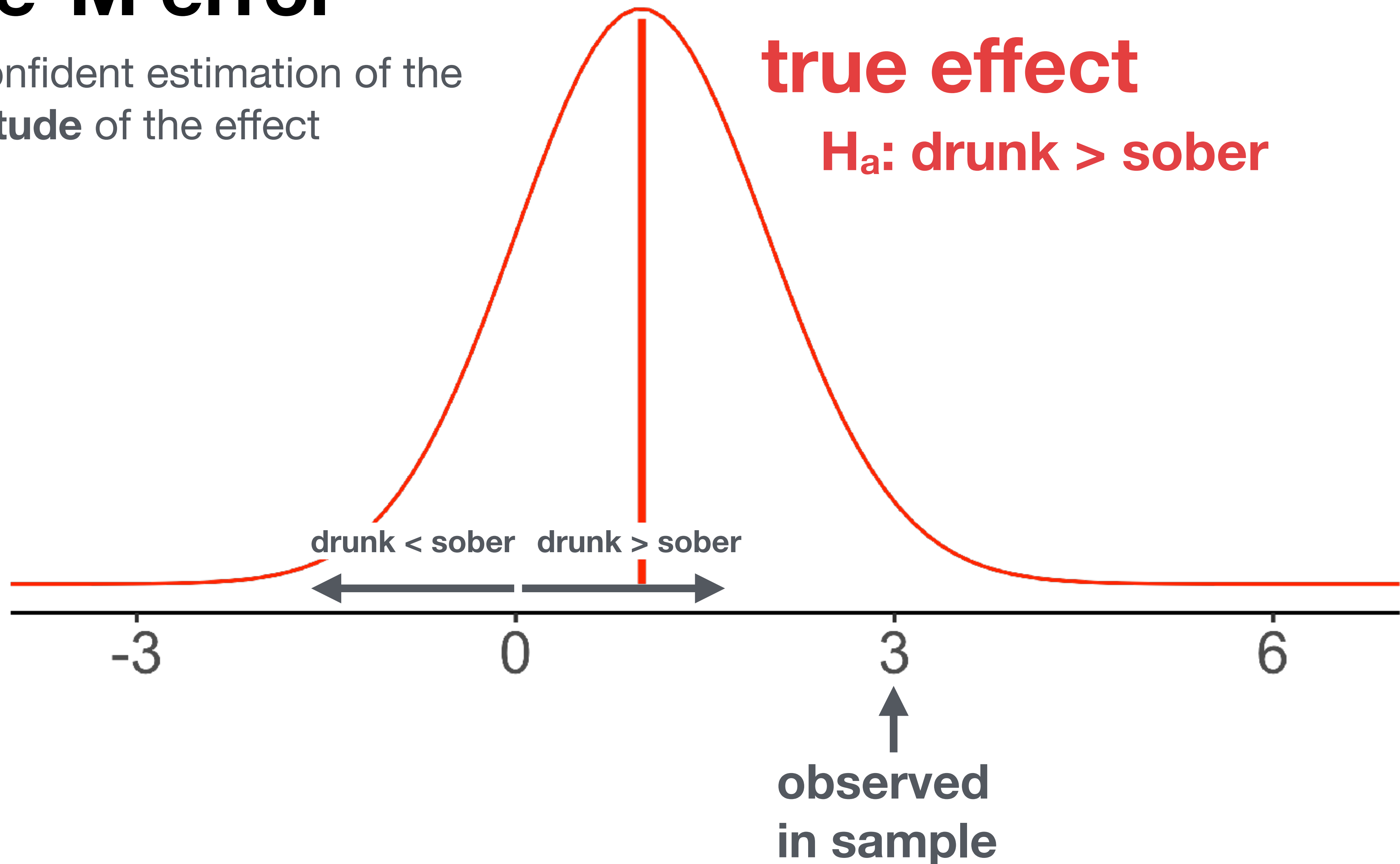
Overconfident estimation of the **magnitude** of the effect

Type-S error

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

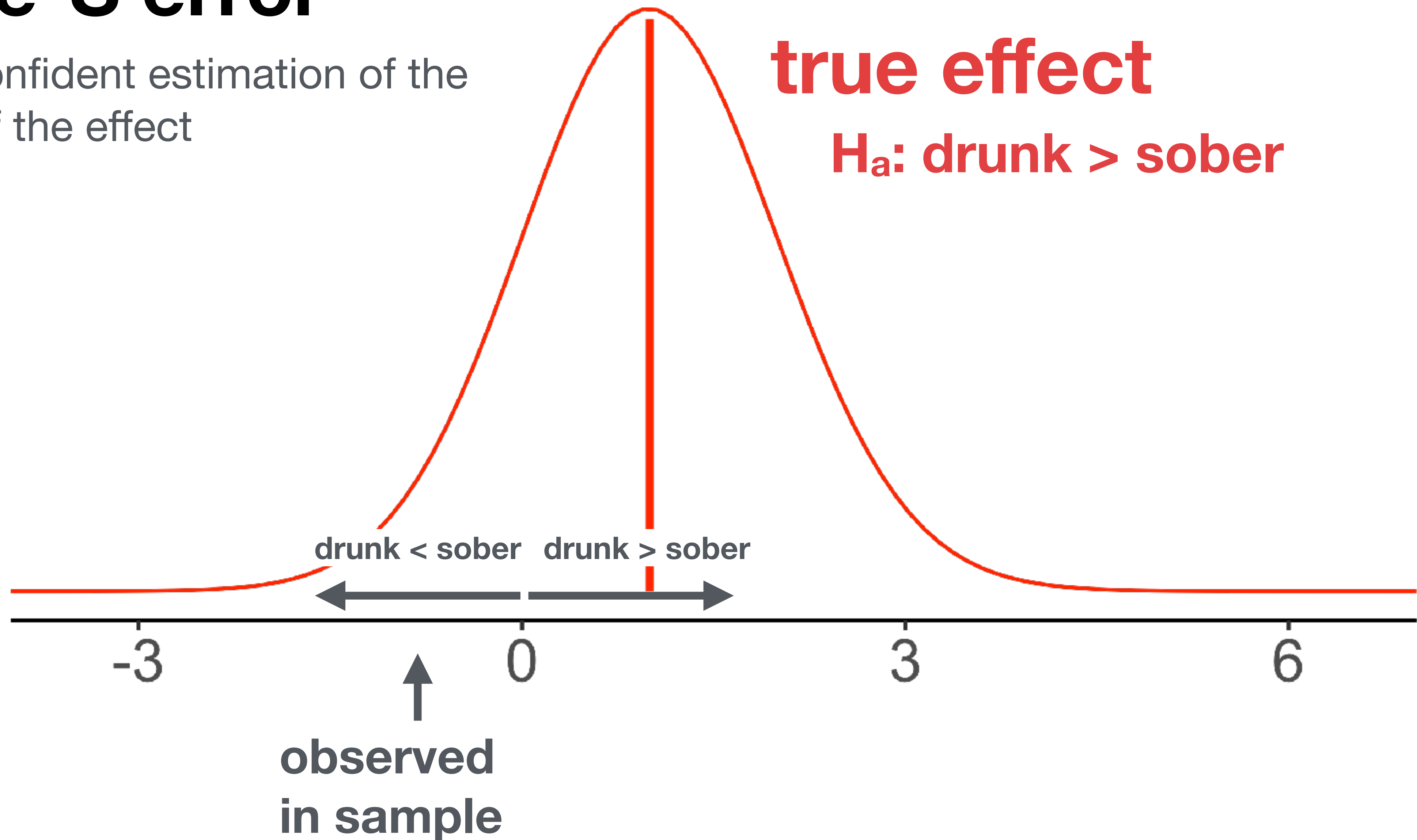
Type-M error

Overconfident estimation of the
magnitude of the effect



Type-S error

Overconfident estimation of the
sign of the effect



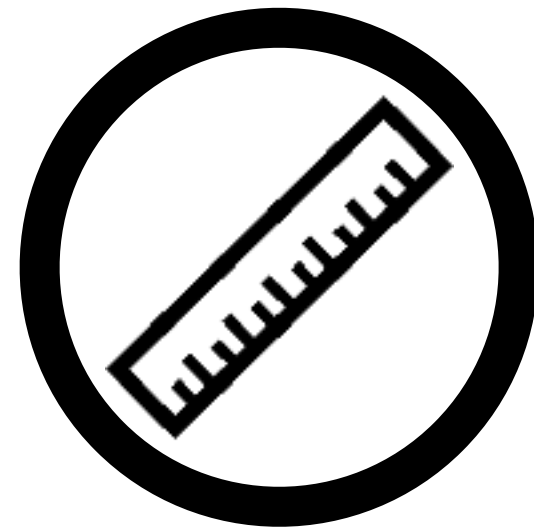


$p < 0.05$

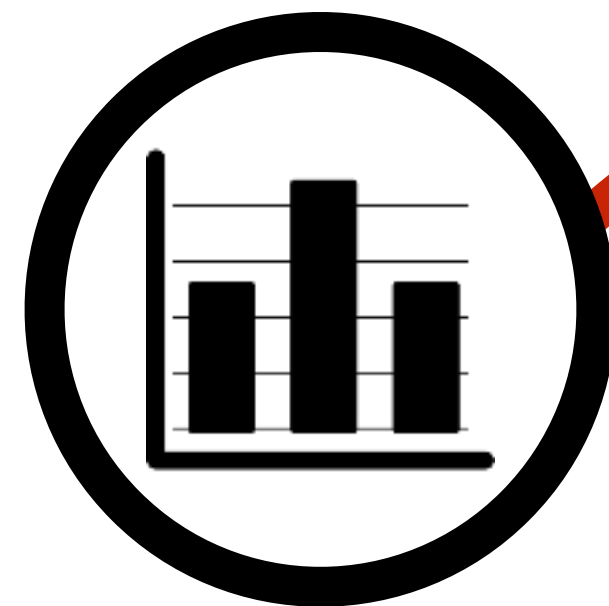


$p < 0.05$





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

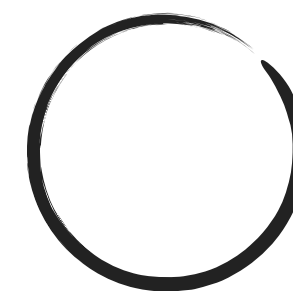
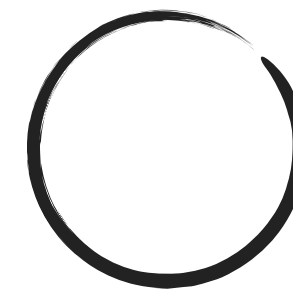
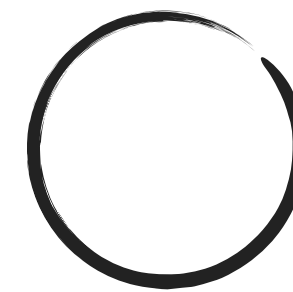
The probability of the
observed result, plus more
extreme results, **if the null
hypothesis was true**



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**



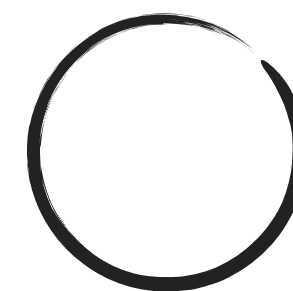
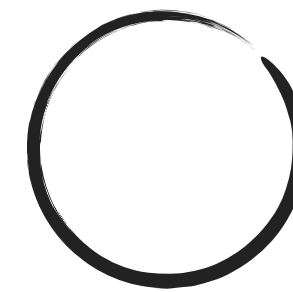
The probability of the null hypothesis



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



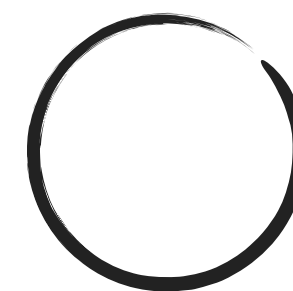
The probability of the alternative hypothesis



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



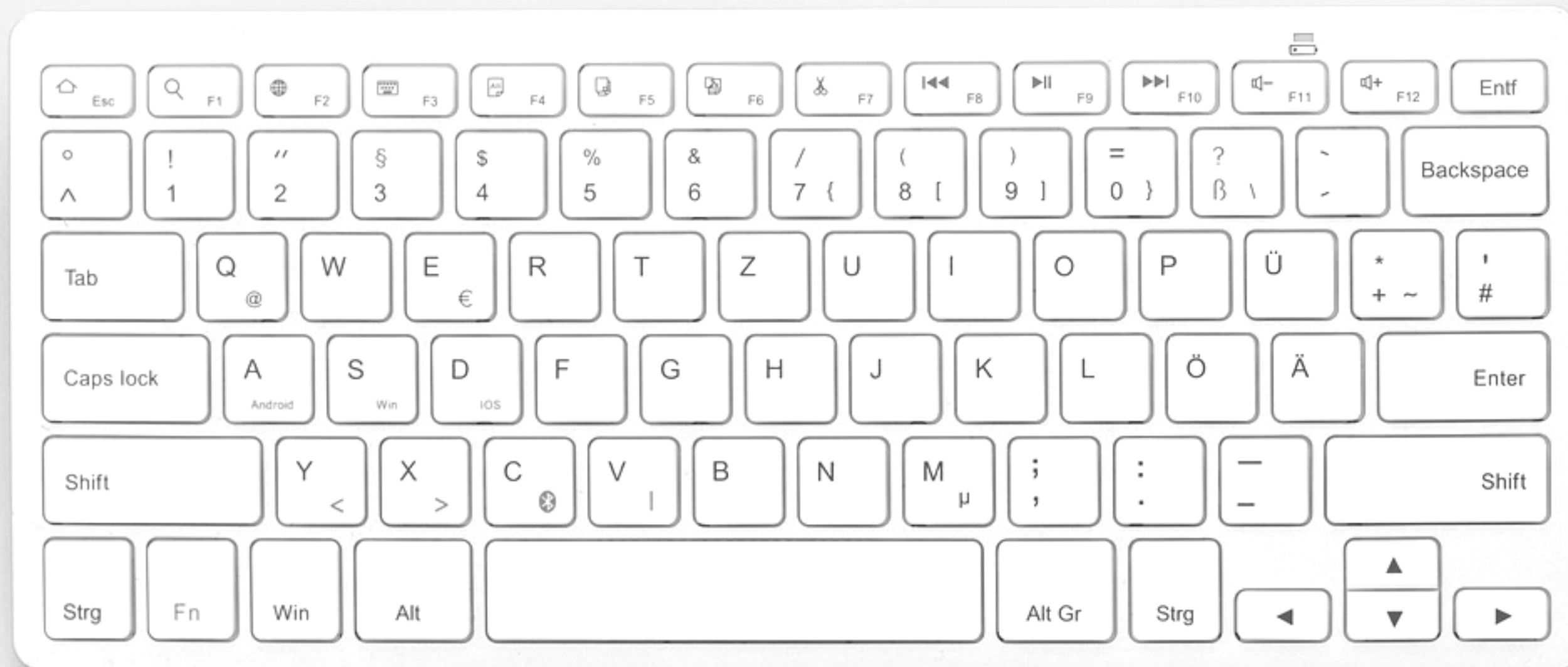
If < 0.05 , we can conclusively answer a scientific question





Unlucky sampling

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





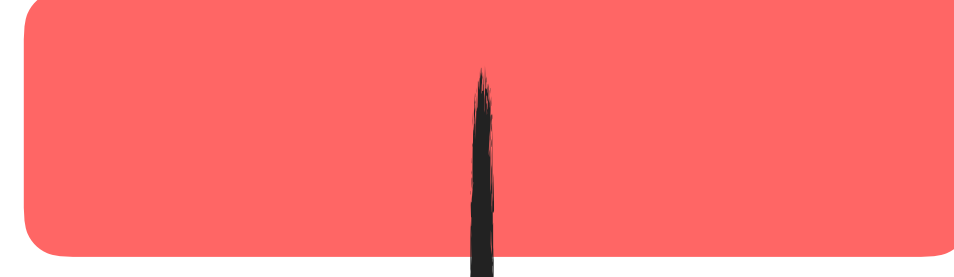
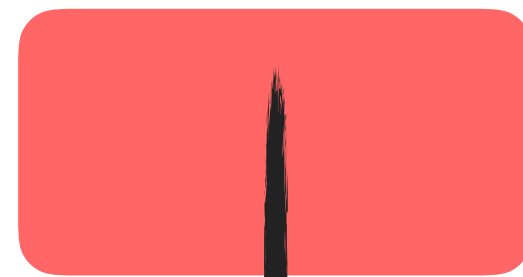
What are your questions?





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

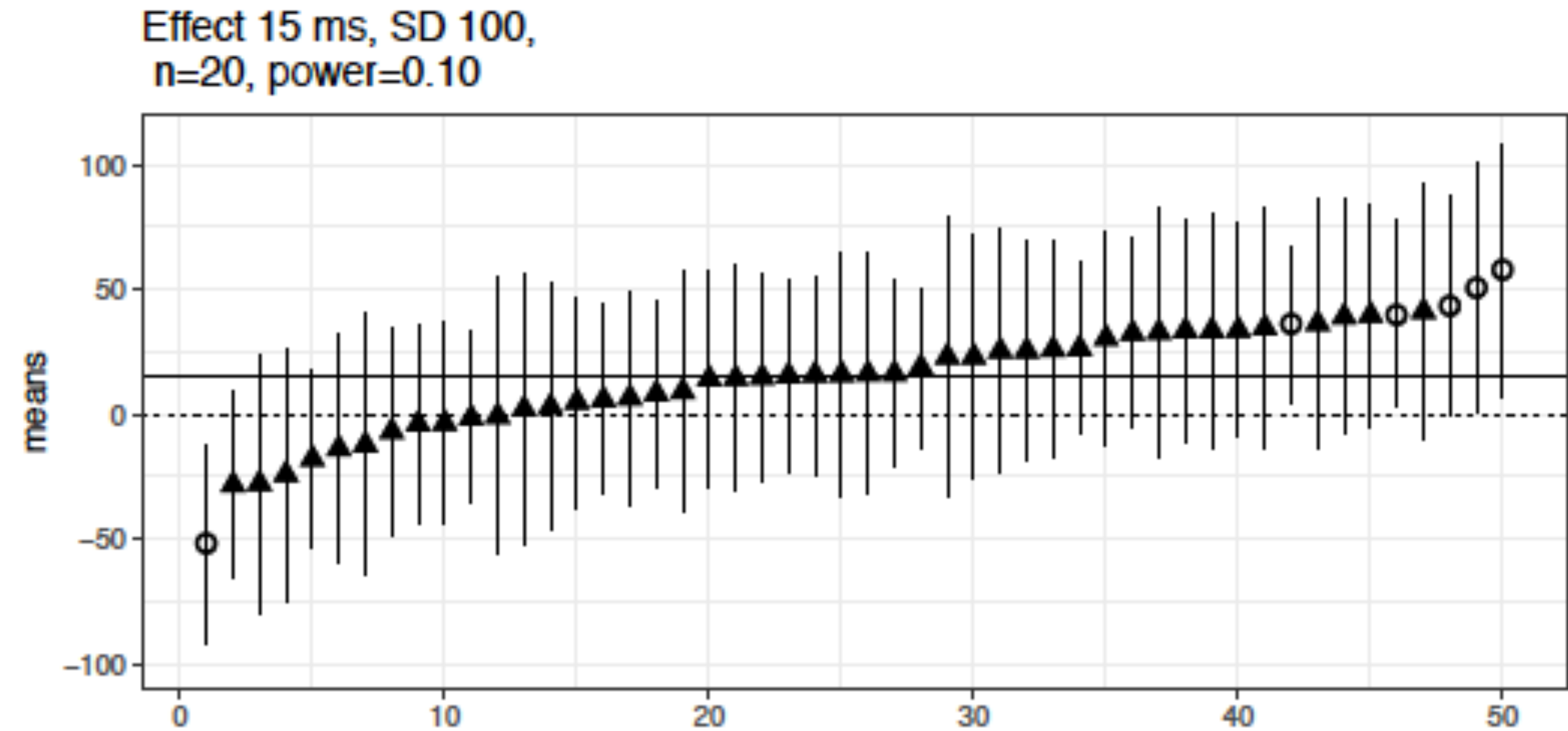
Probability of correctly failing to reject the null hypothesis

Probability of erroneously failing to reject the null

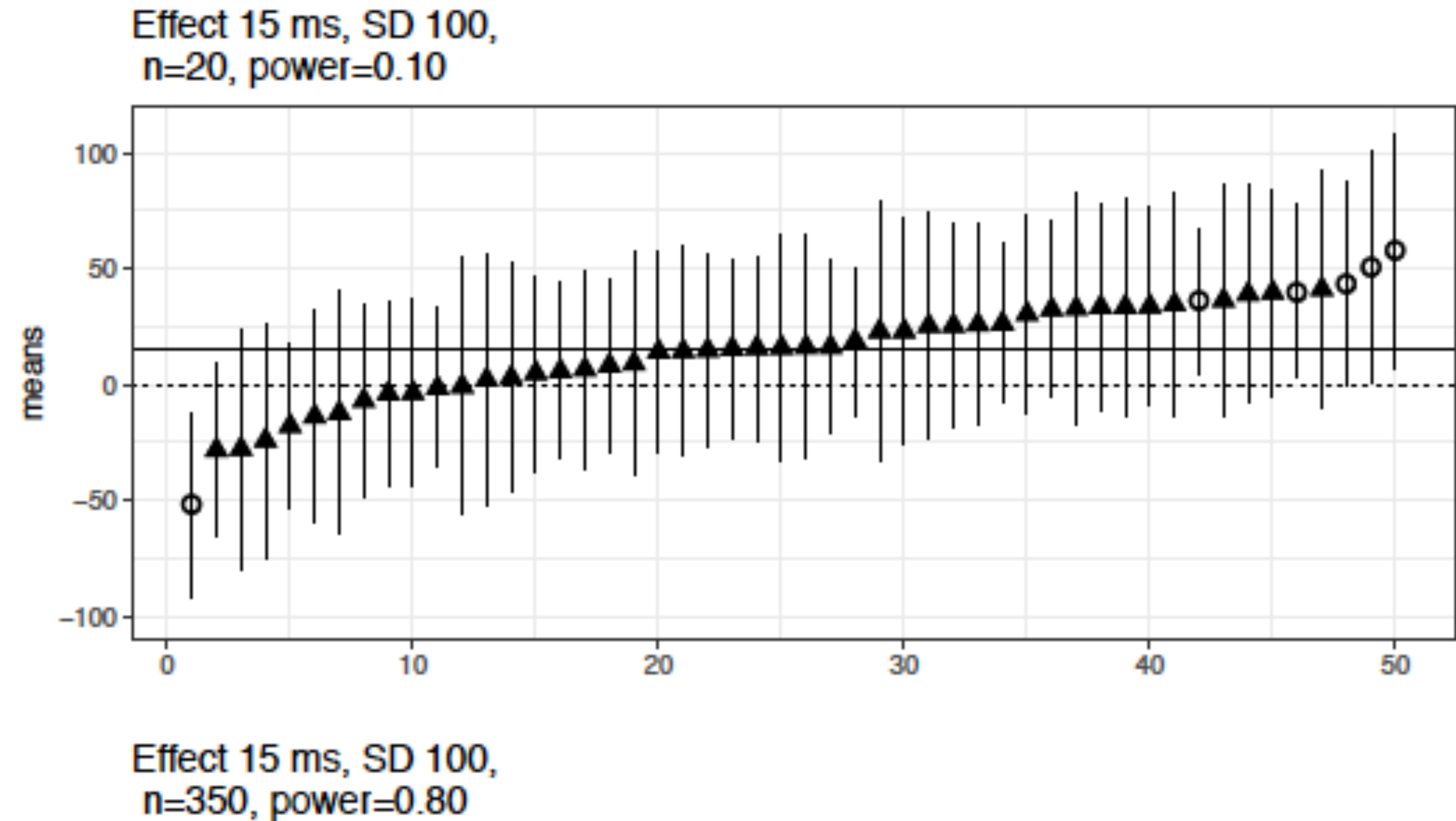
The diagram illustrates the calculation of a t-value. A blue arrow points down to a blue box containing the number 4, which is labeled "t value". This is followed by an equals sign. To the right of the equals sign is a fraction. The numerator of the fraction is a red box containing "4 - 2", which is labeled "difference between groups" in red text. The denominator is a green box containing "24", which is labeled "sample size" in green text. A green arrow points down from the "sample size" label. A green bracket on the left side of the fraction is labeled "standard error (SE)" in green text. The fraction is also labeled "noise" in green text. The entire formula is presented as:

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

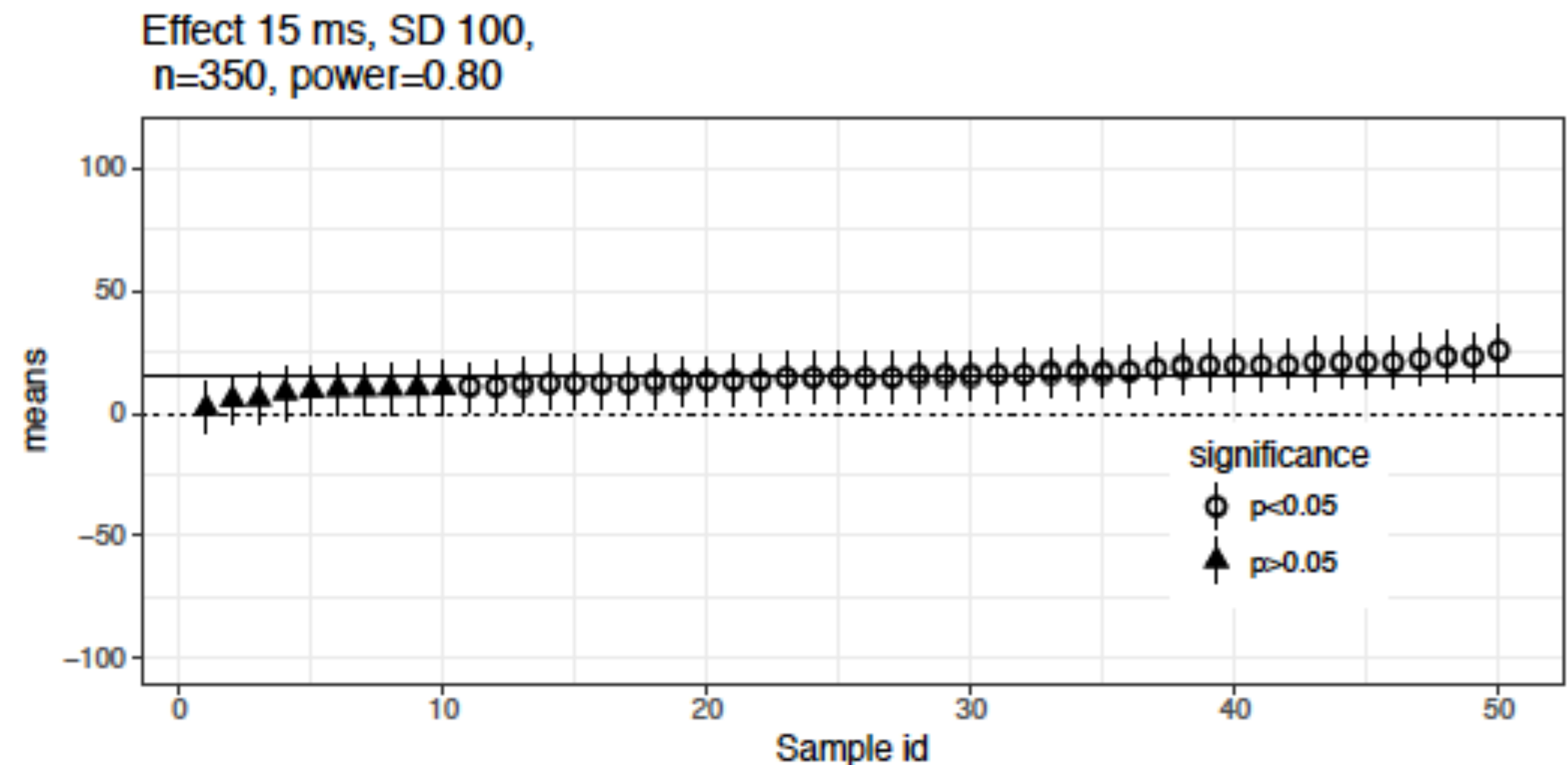
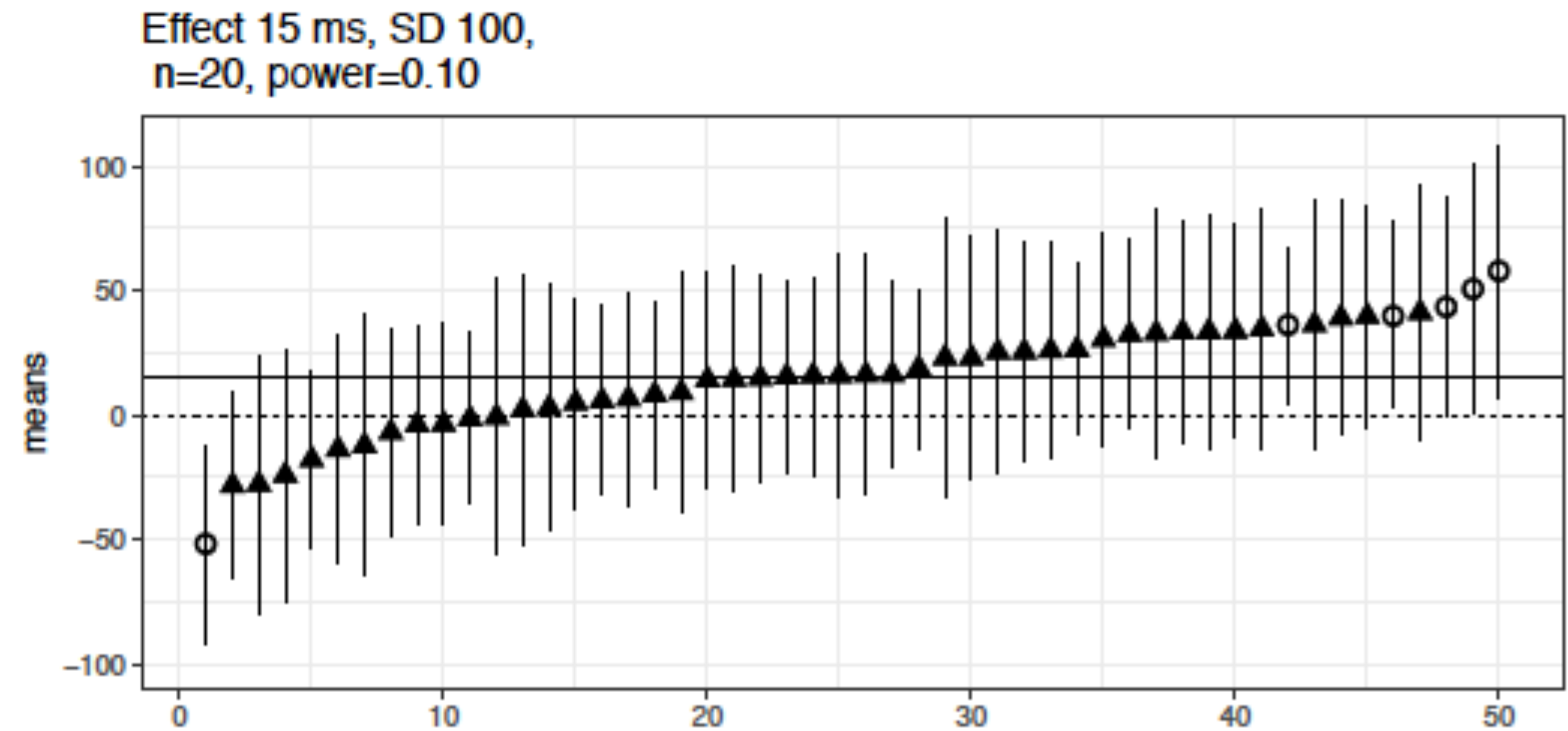
Small
samples size
leads to increased
Type M &
Type S errors



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

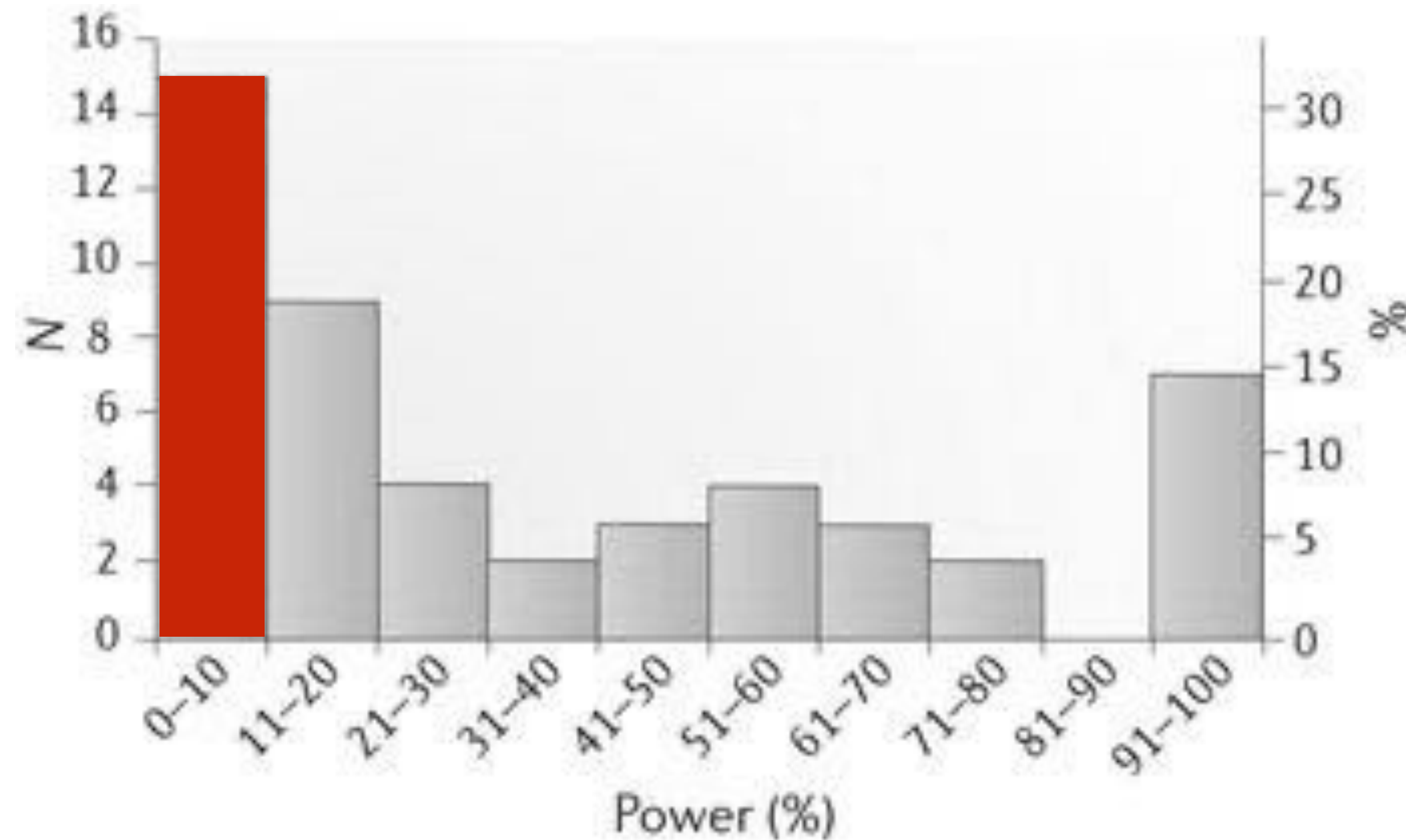


**Small
samples size
leads to increased
Type M &
Type S errors**



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?

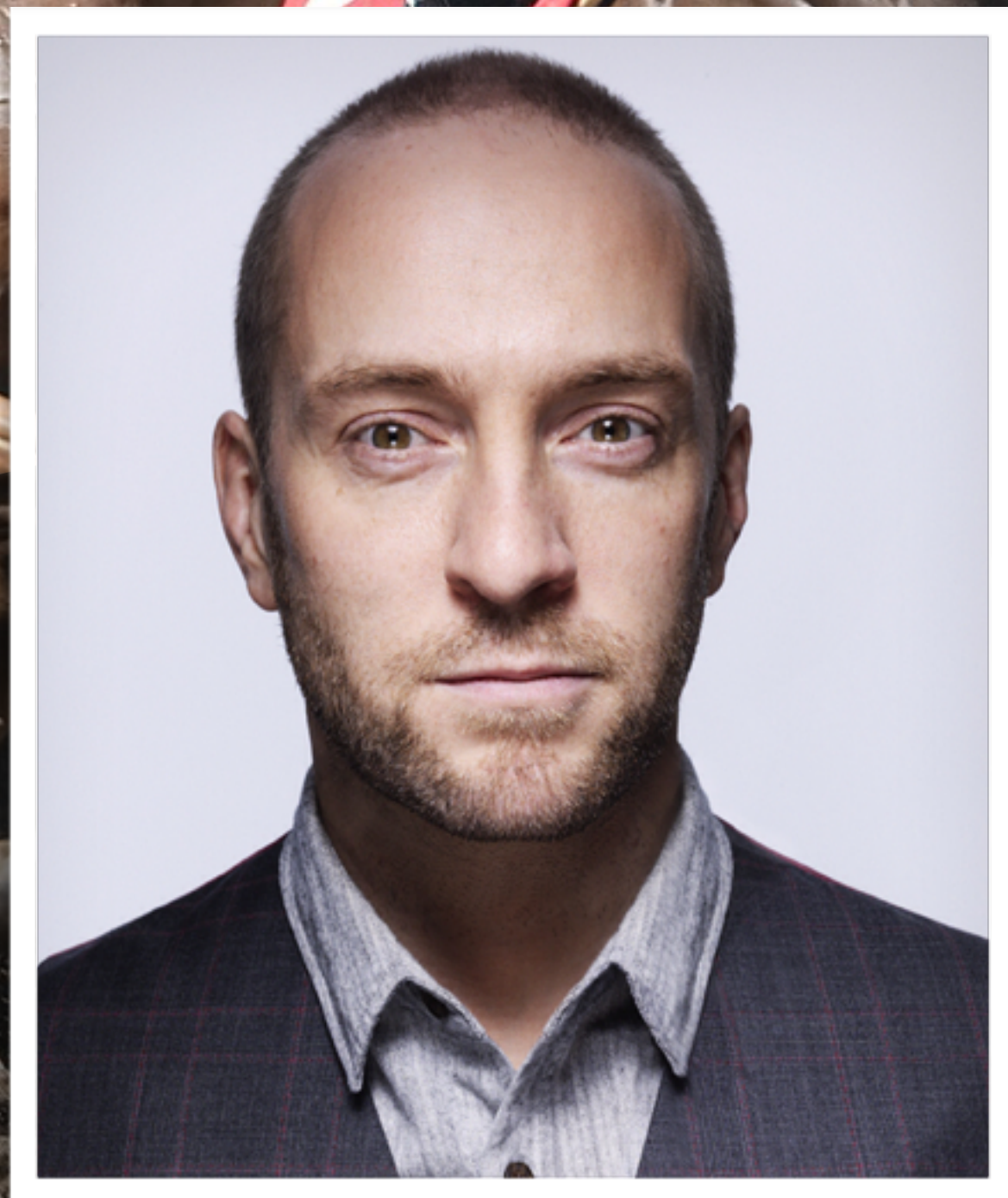
We make statistical **errors**

Our studies have not enough **power**

We explore **researcher degrees of freedom**



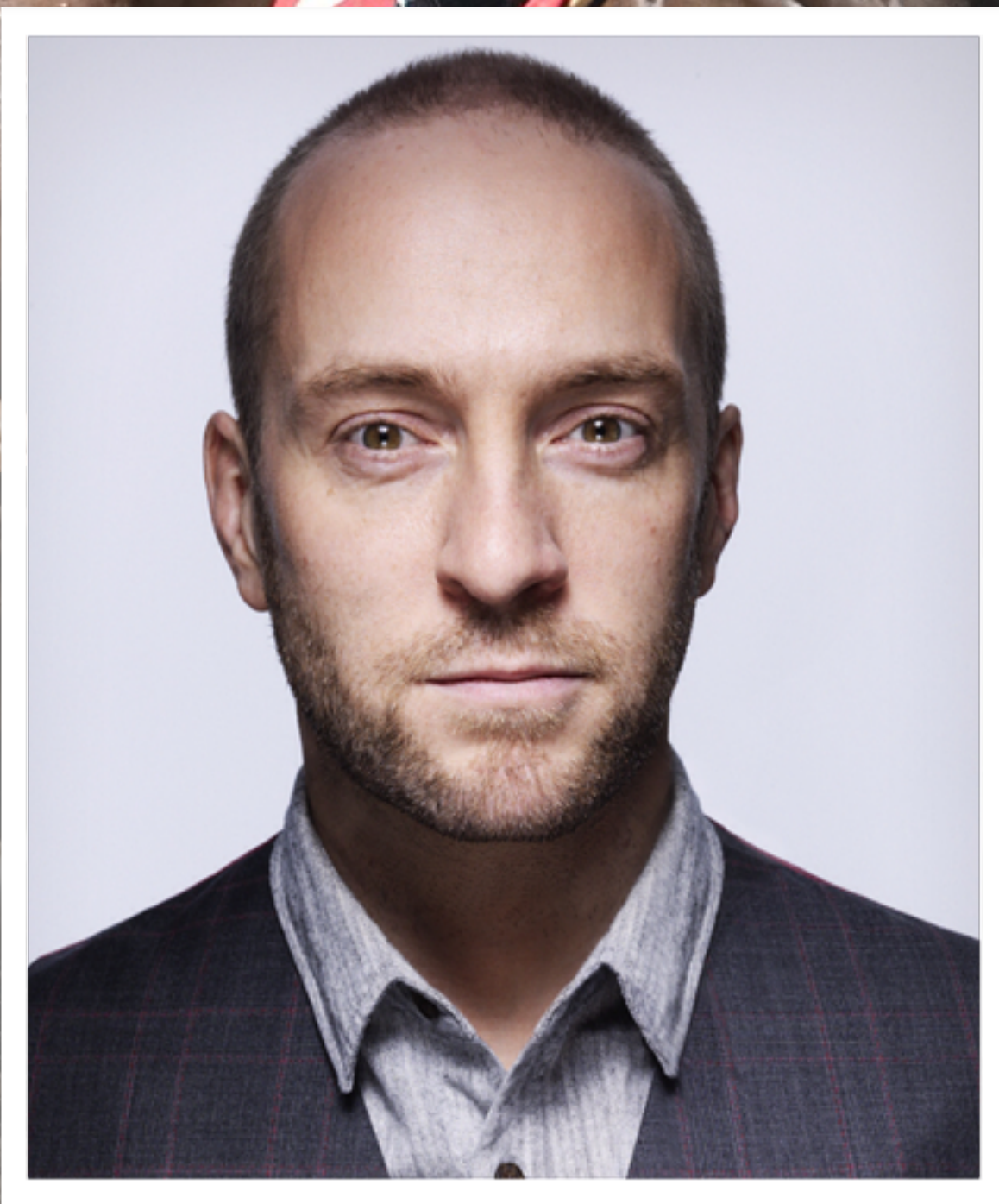
The winning streak



Derren Brown



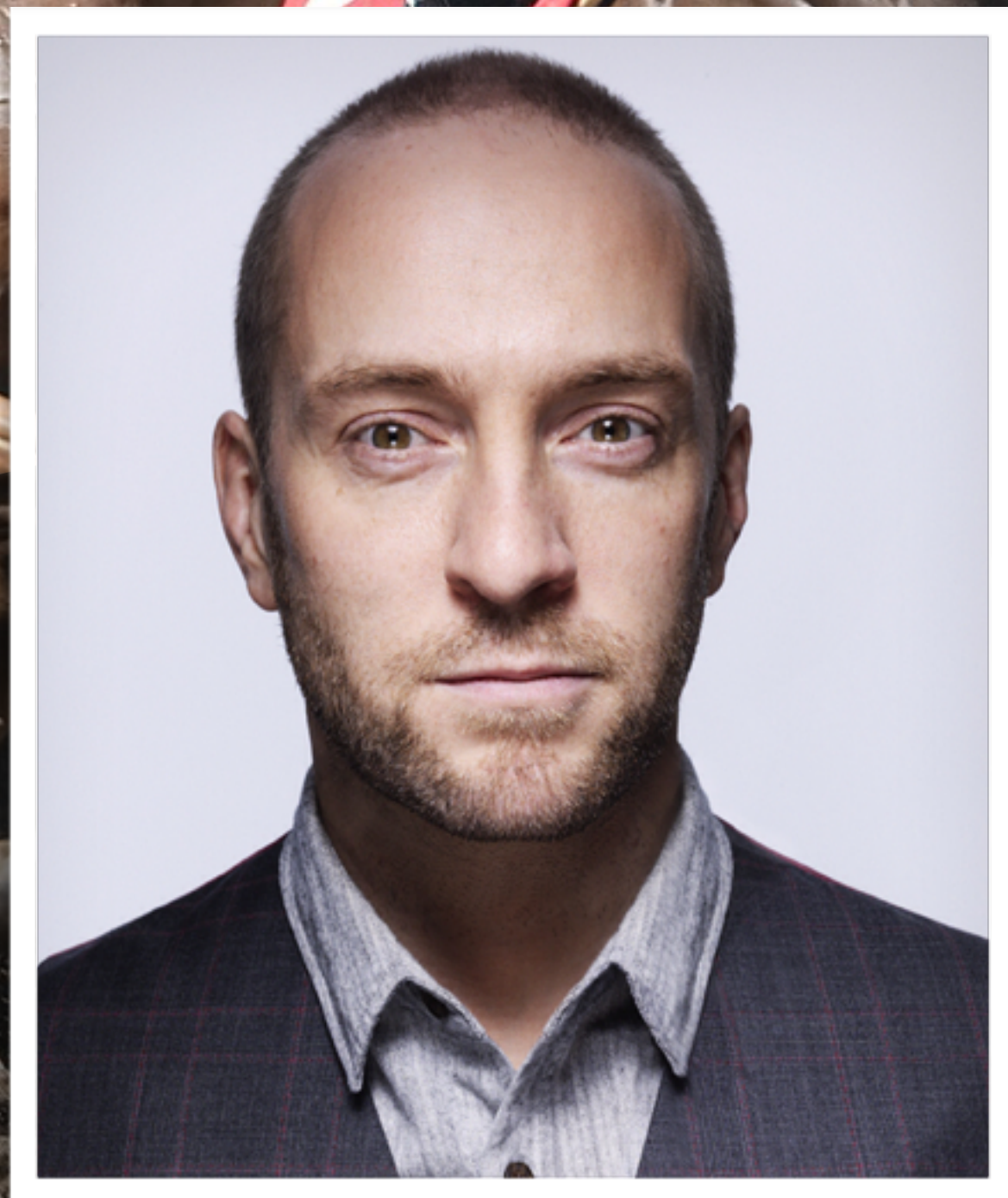
The winning streak



Derren Brown



The winning streak

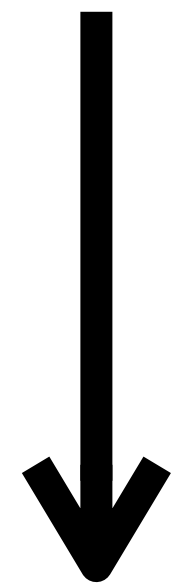


Derren Brown

Multiple testing

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

H_0 : Berlin = Osnabrück



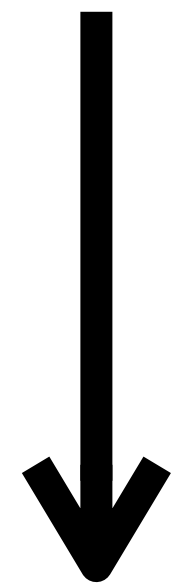
More V-necks
 $p = 0.6$



Multiple testing

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

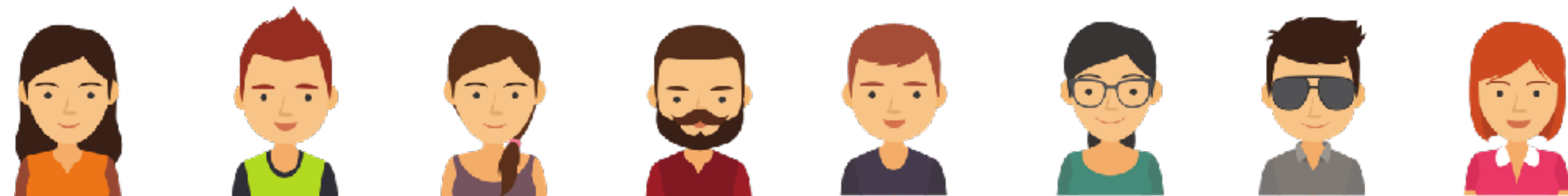
H_0 : Berlin = Osnabrück



More V-necks
 $p = 0.6$

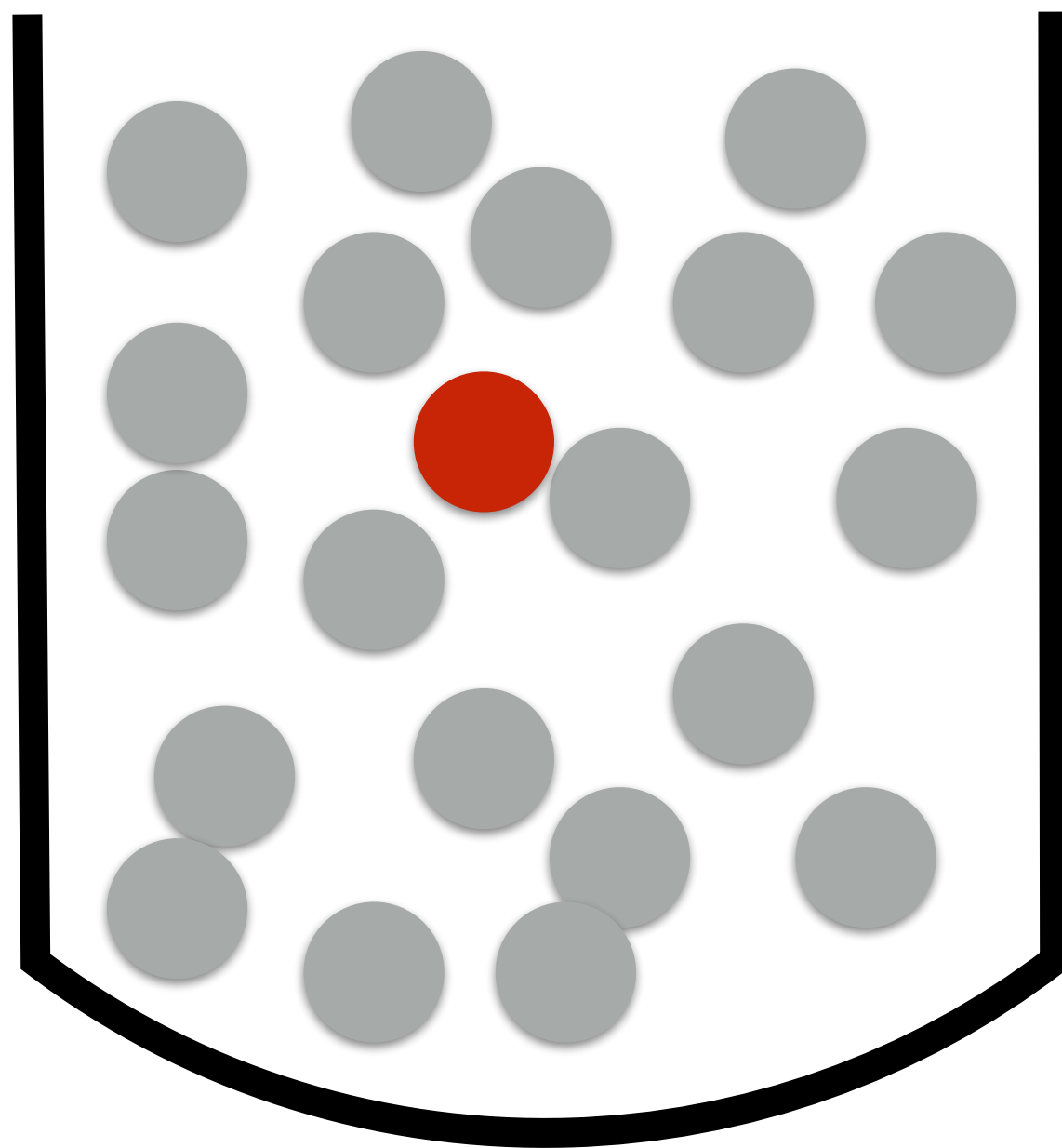


More colorful
 $p = 0.04$



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

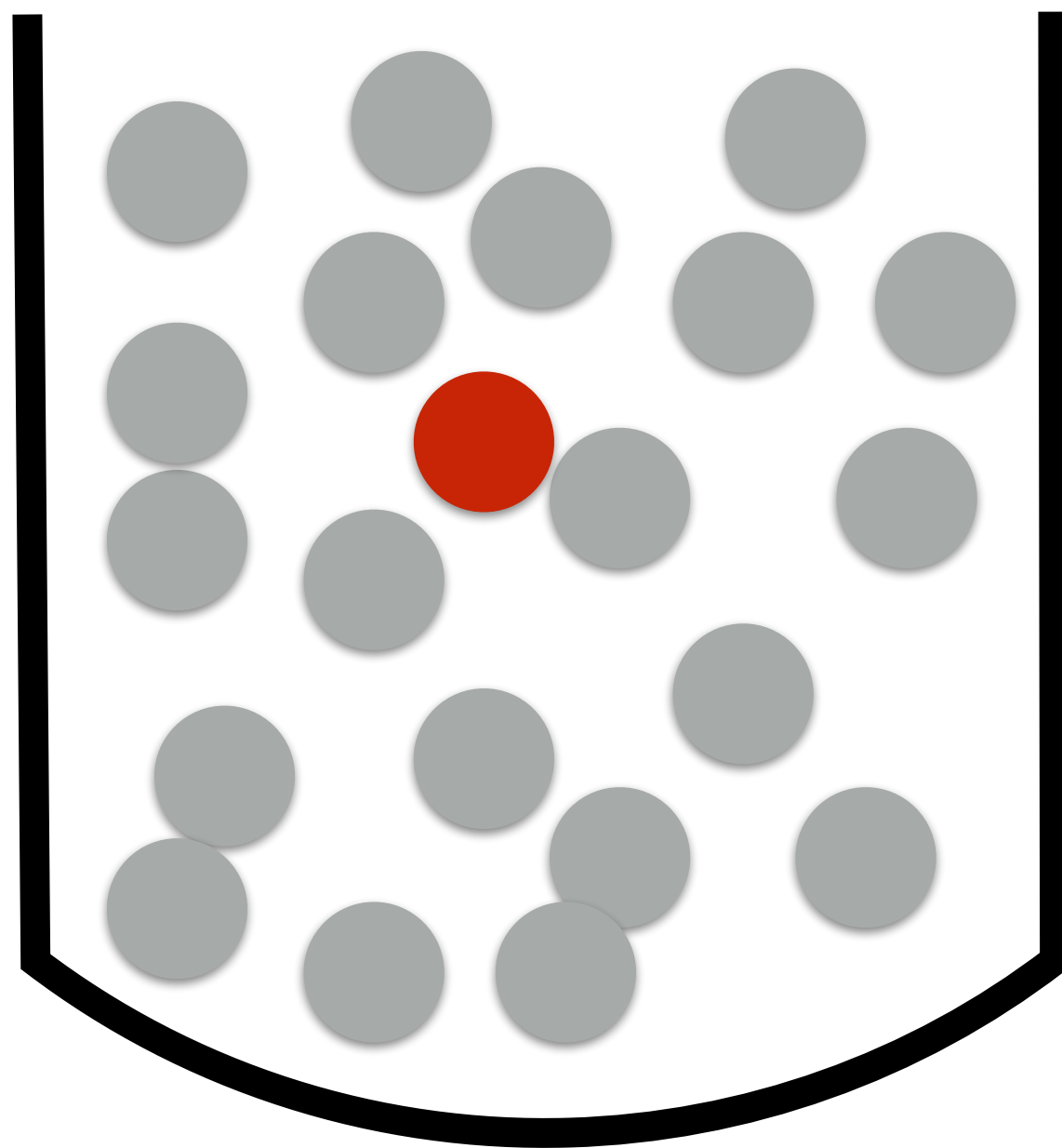
0.05



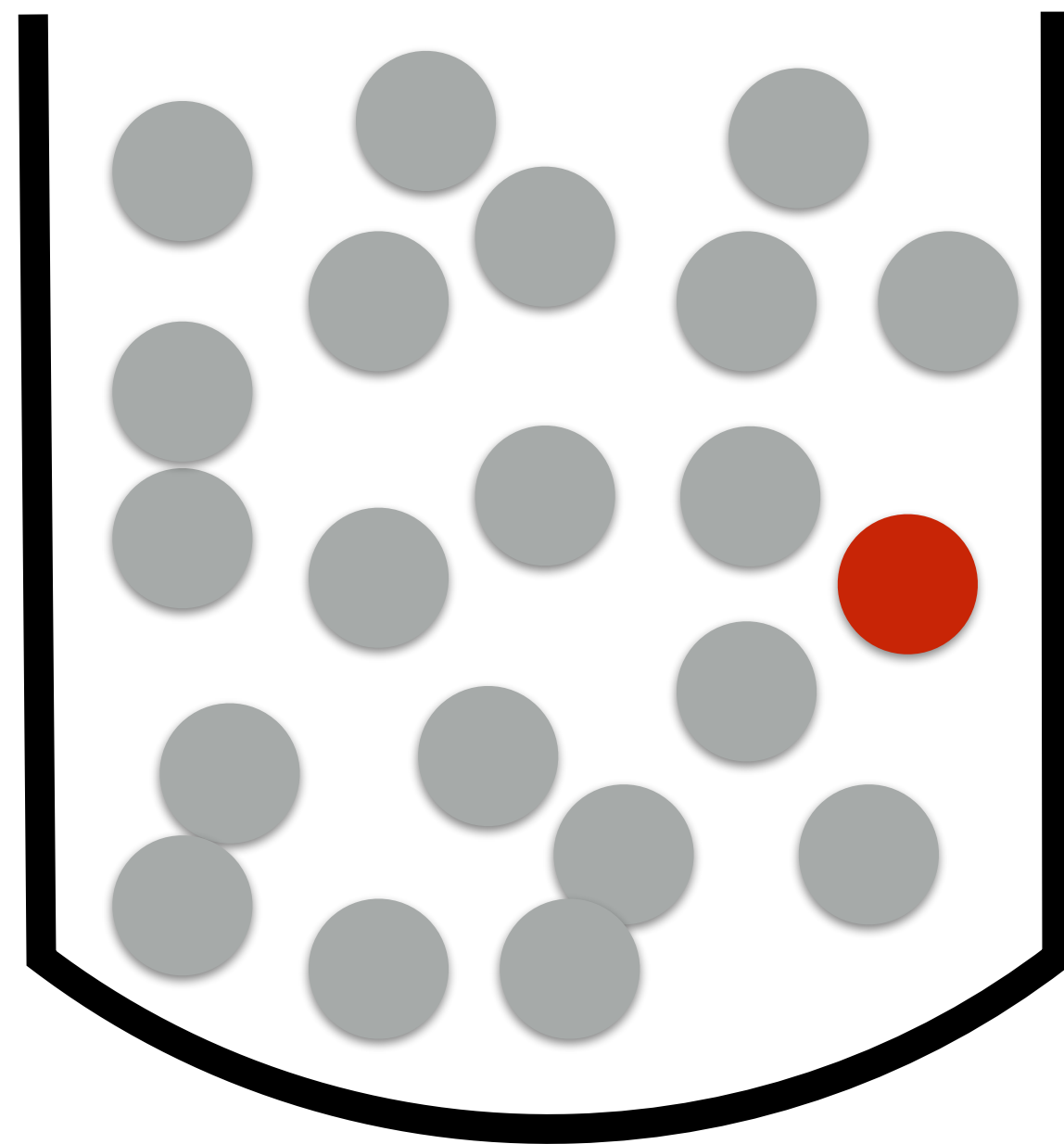
● x 19 ● x 1

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

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



● x 19 ● x 1



● x 19 ● x 1

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

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

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

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>



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

Researcher degrees of freedom in phonetic research

Timo B. Roettger

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

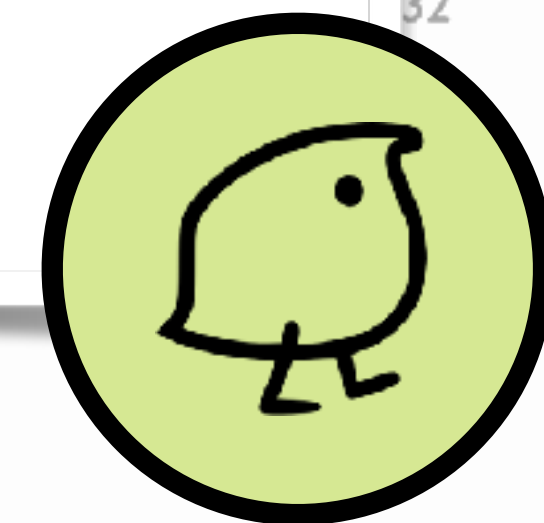
False-Positives
Flexibility
Allows P

can be a problem,
and the research

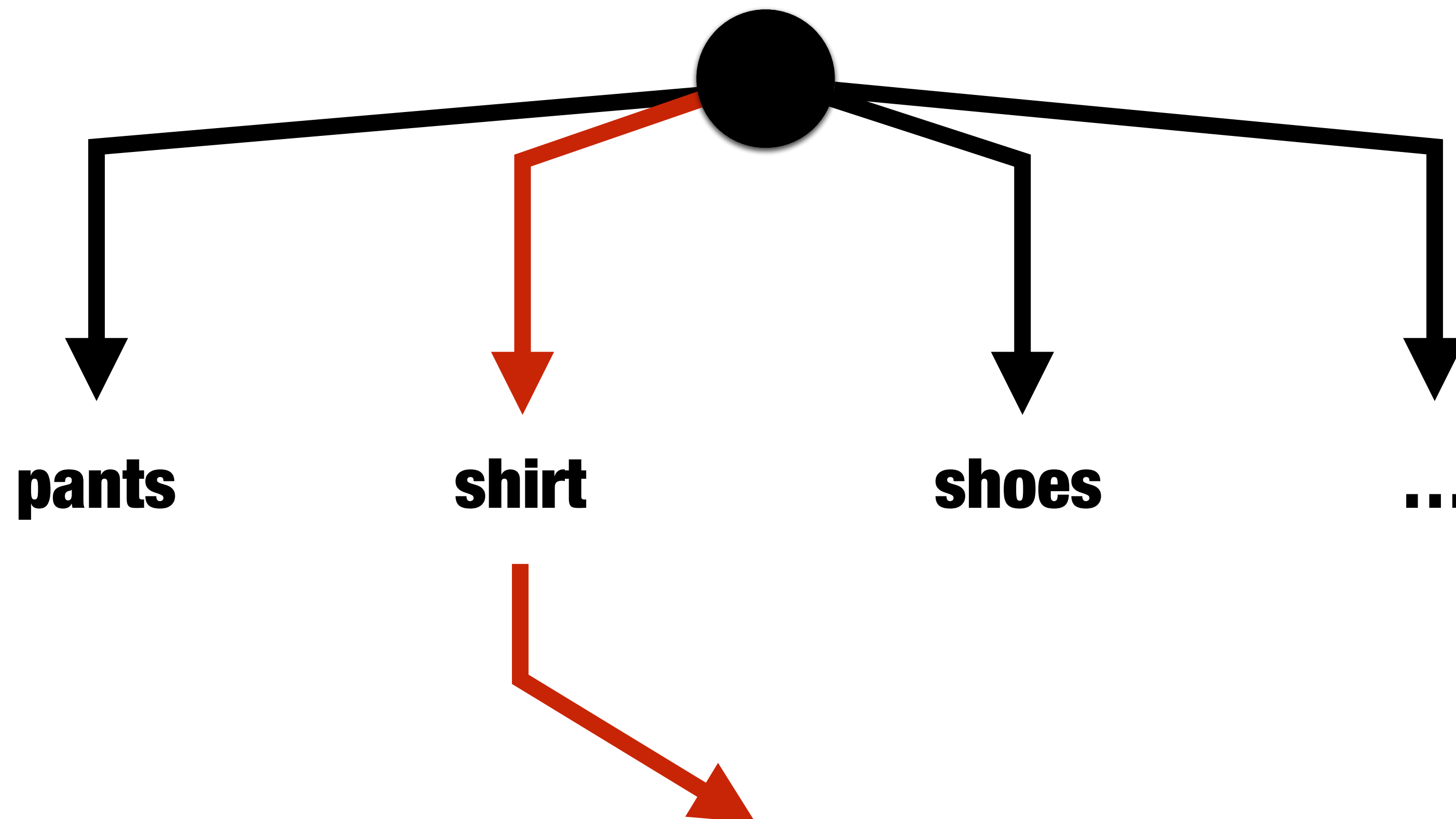
ns.nav
32

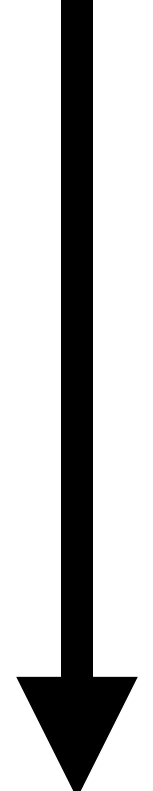
Joseph P. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

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

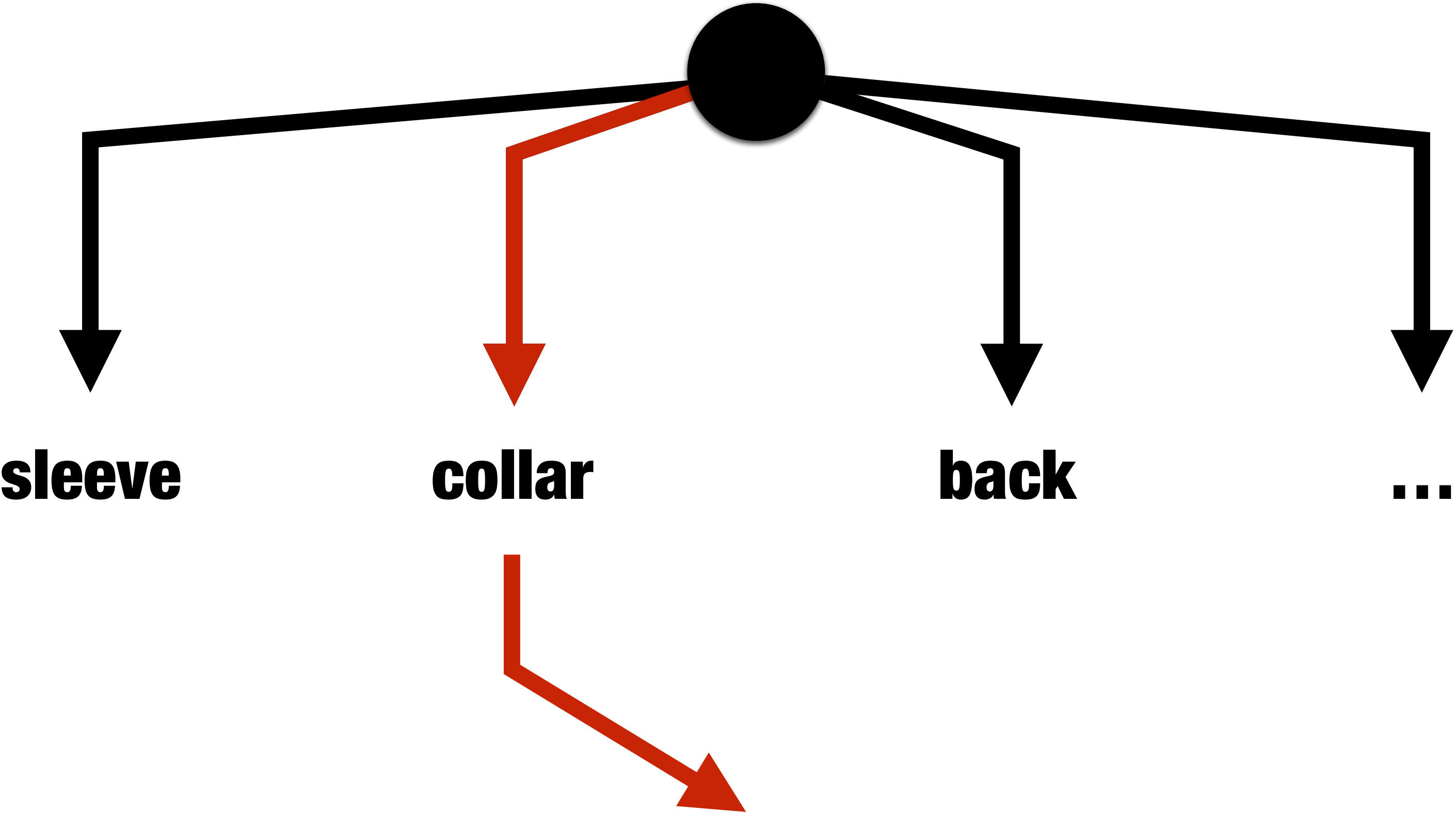


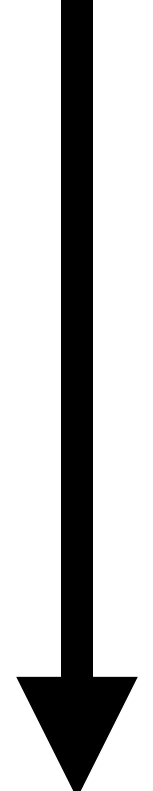
choose a piece of clothing



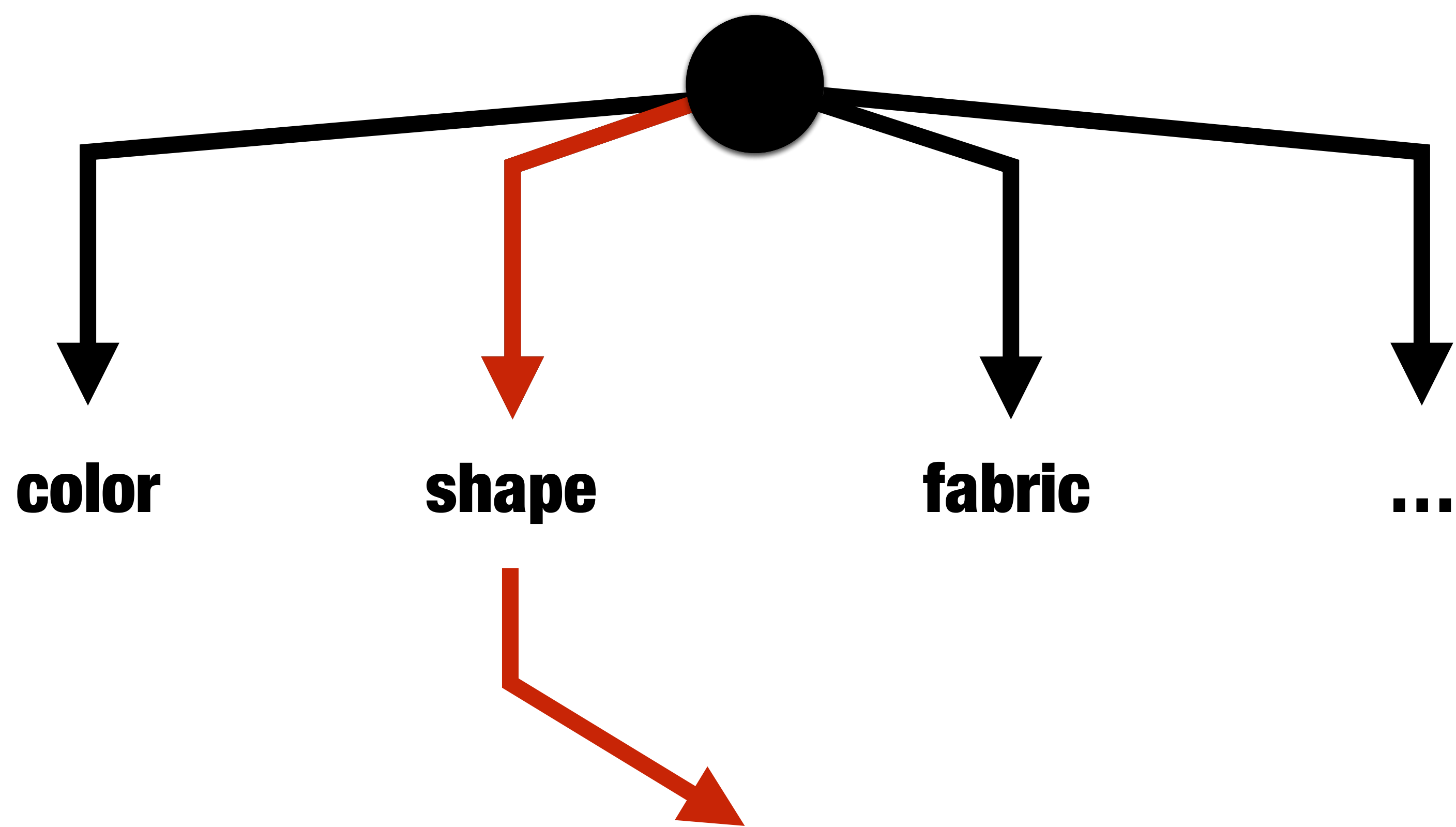


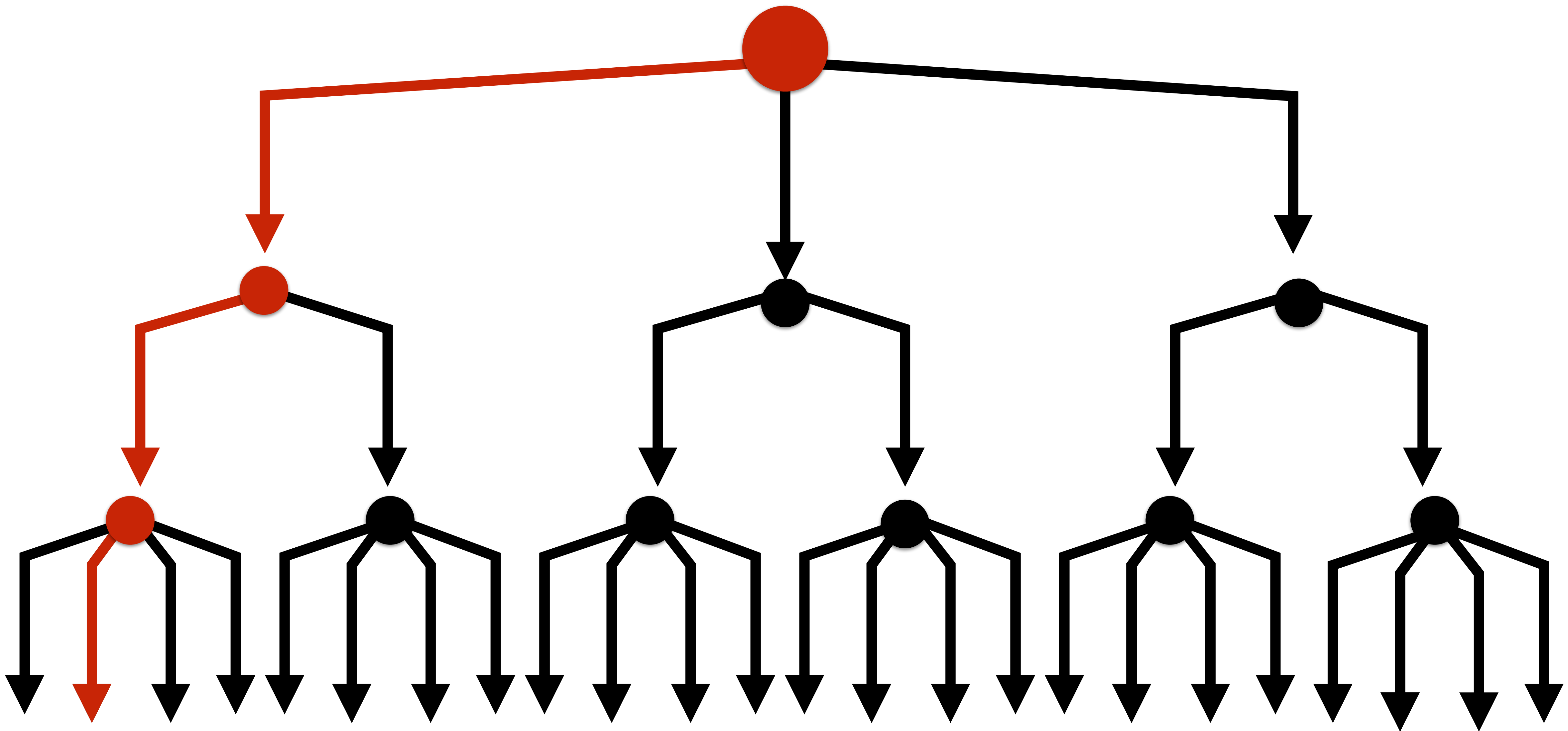
choose part of clothing

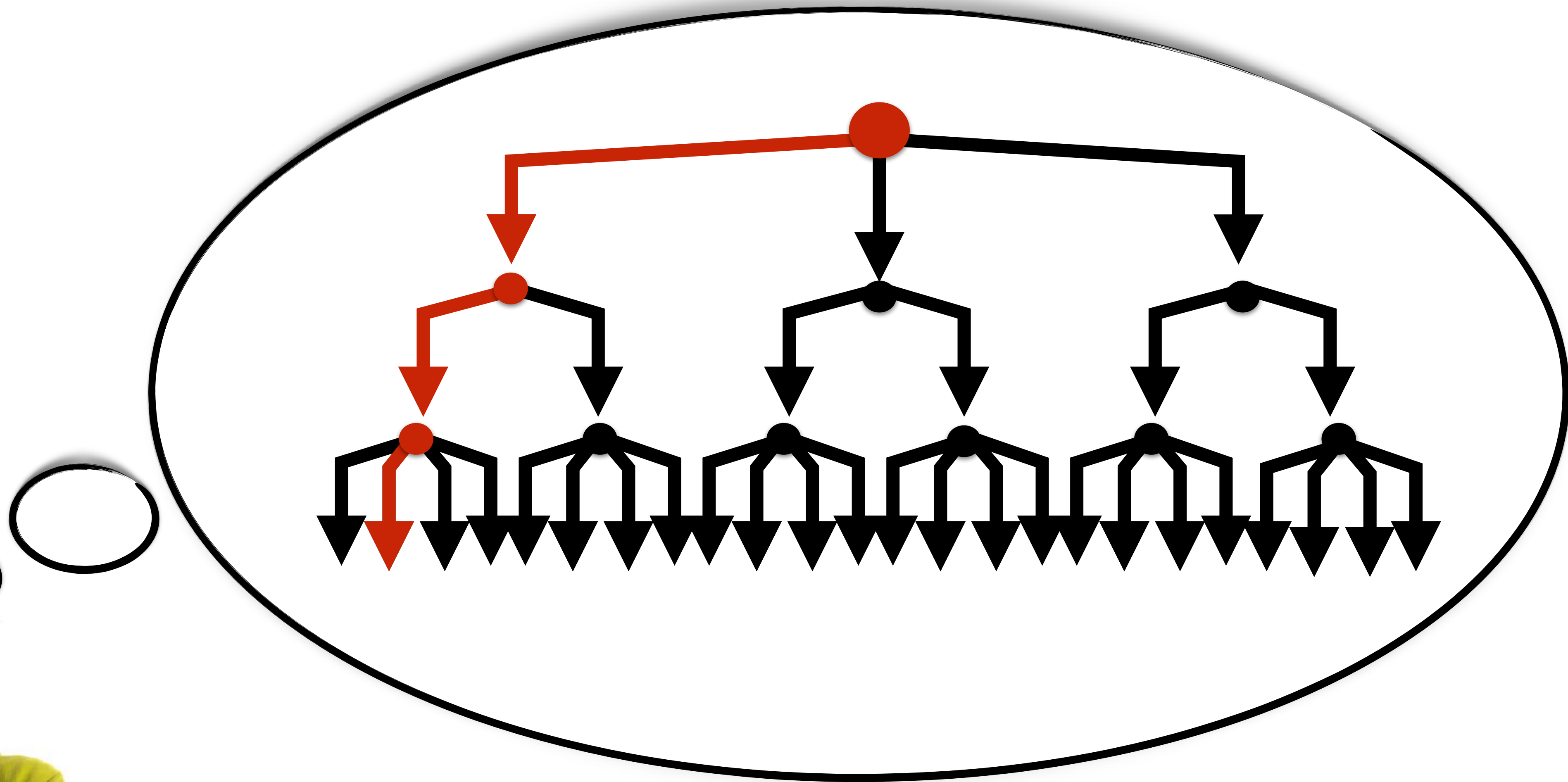


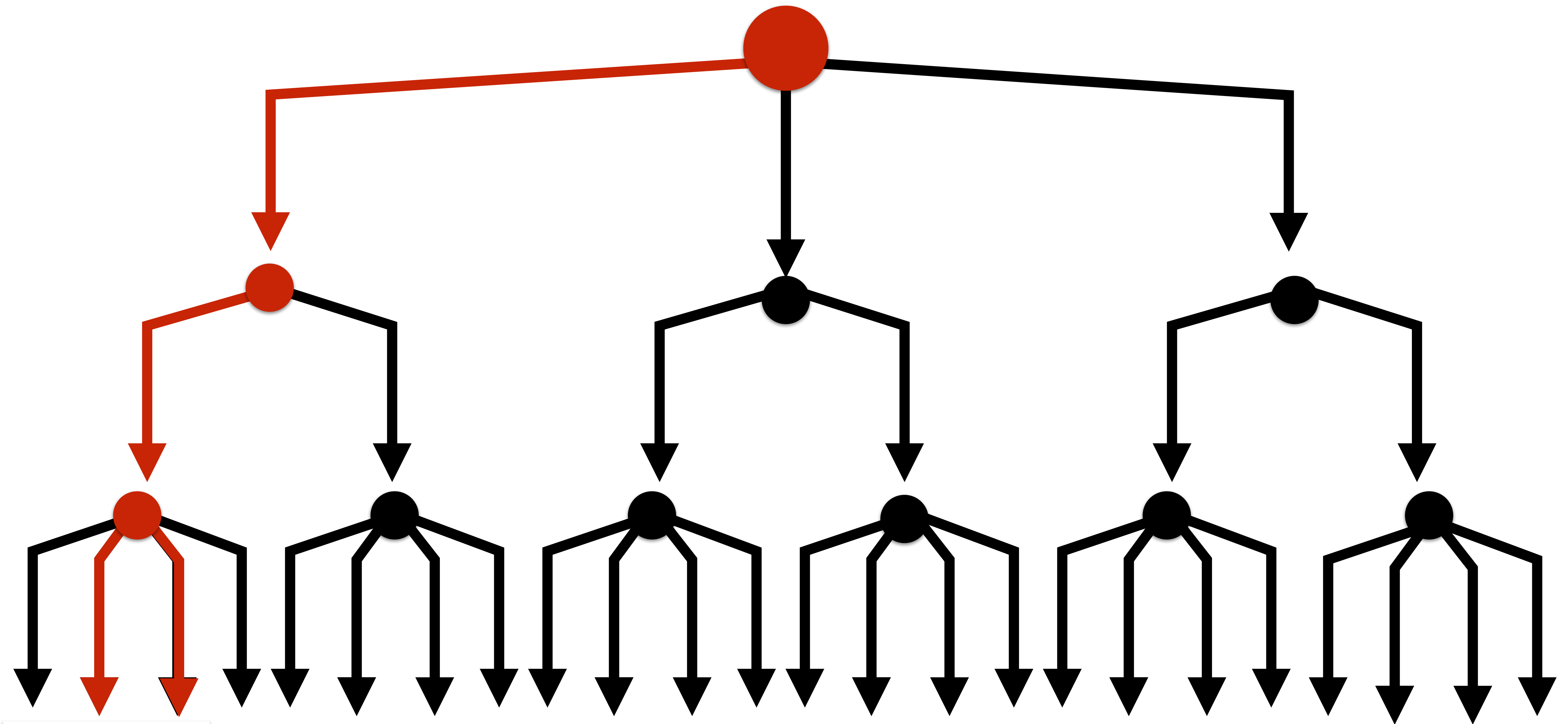


choose concrete operationalization



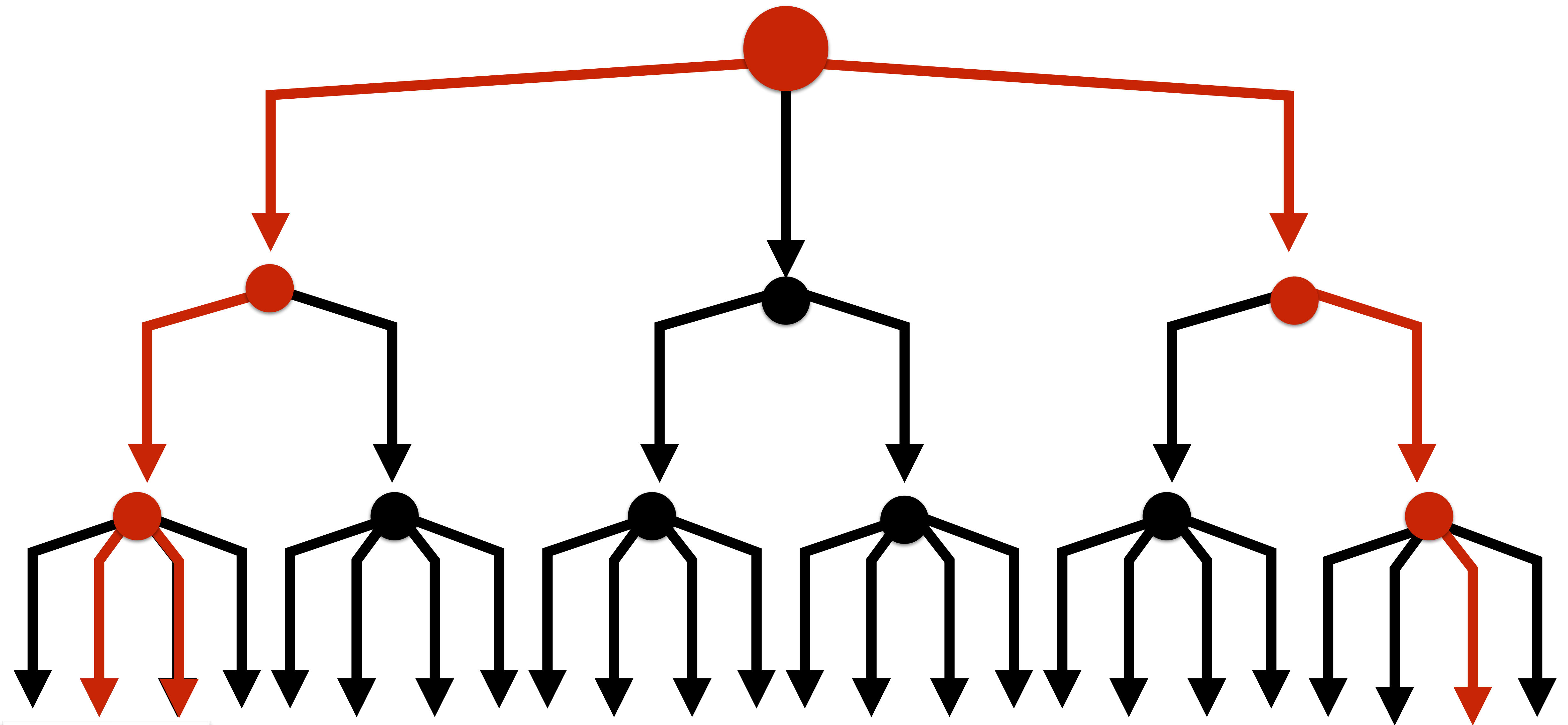






$p = 0.3$

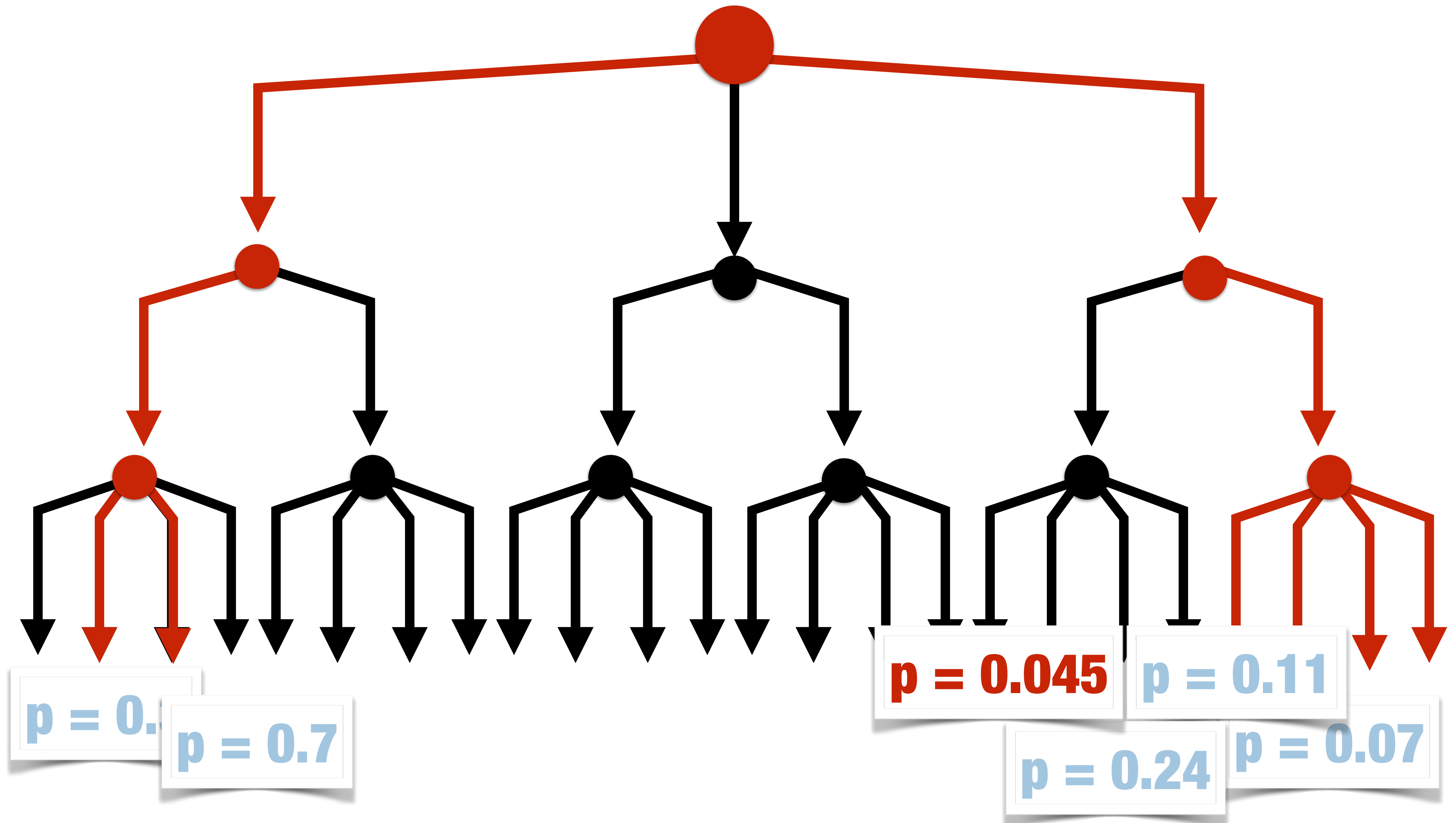
$p = 0.7$

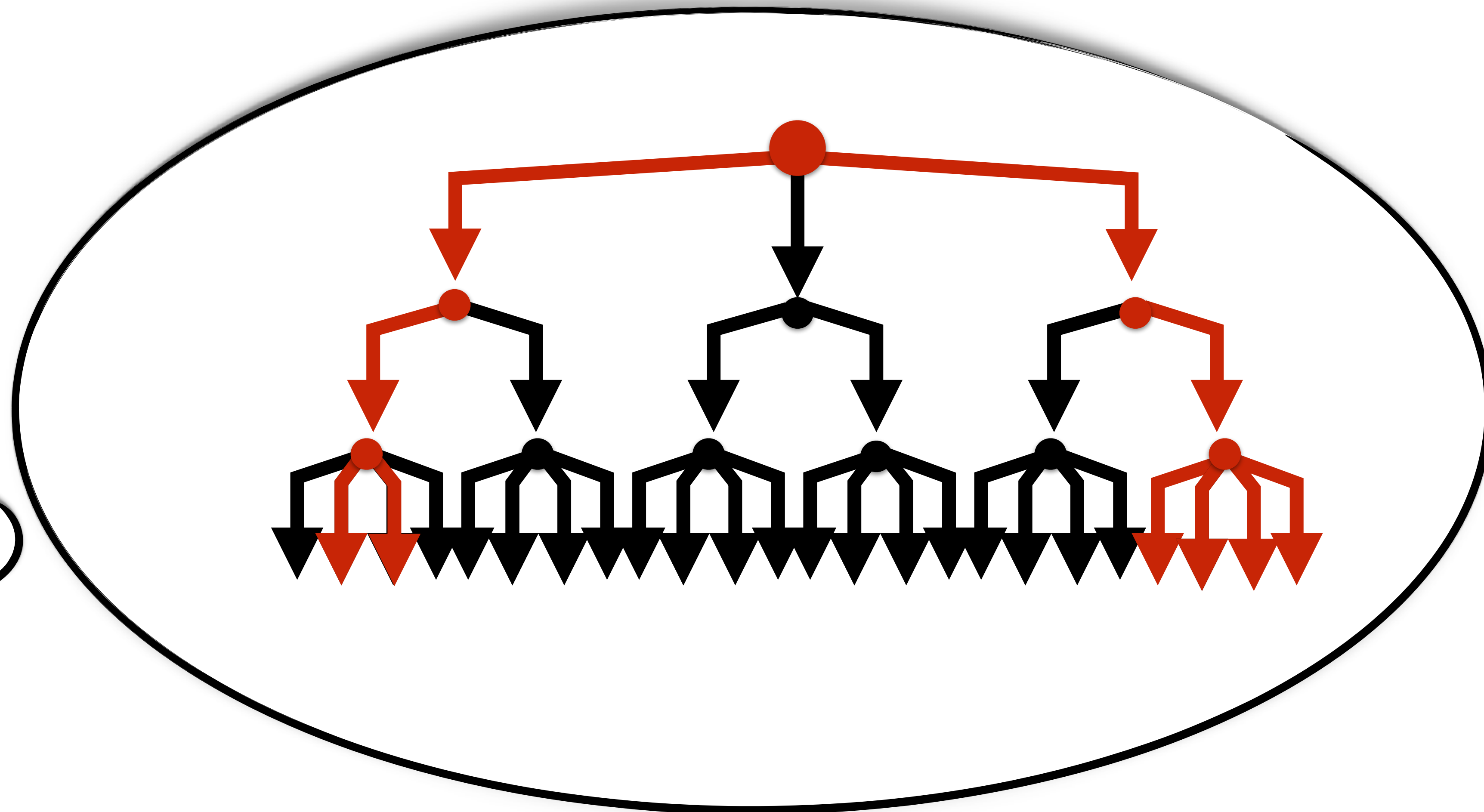
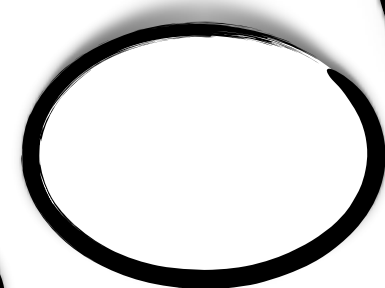


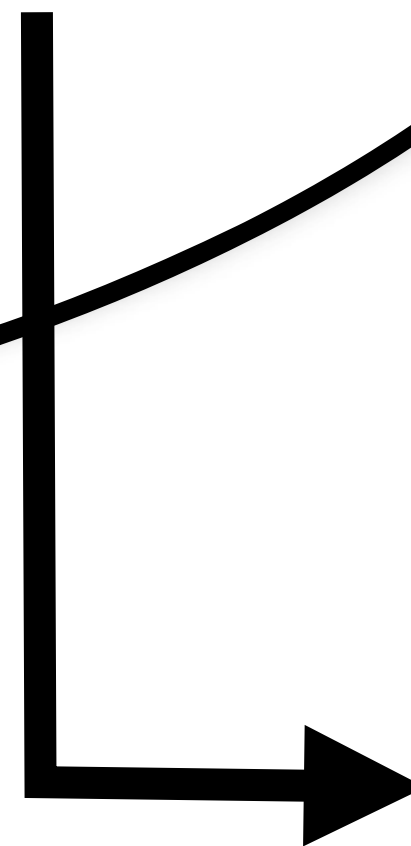
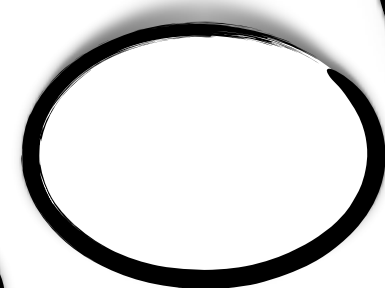
$p = 0.7$

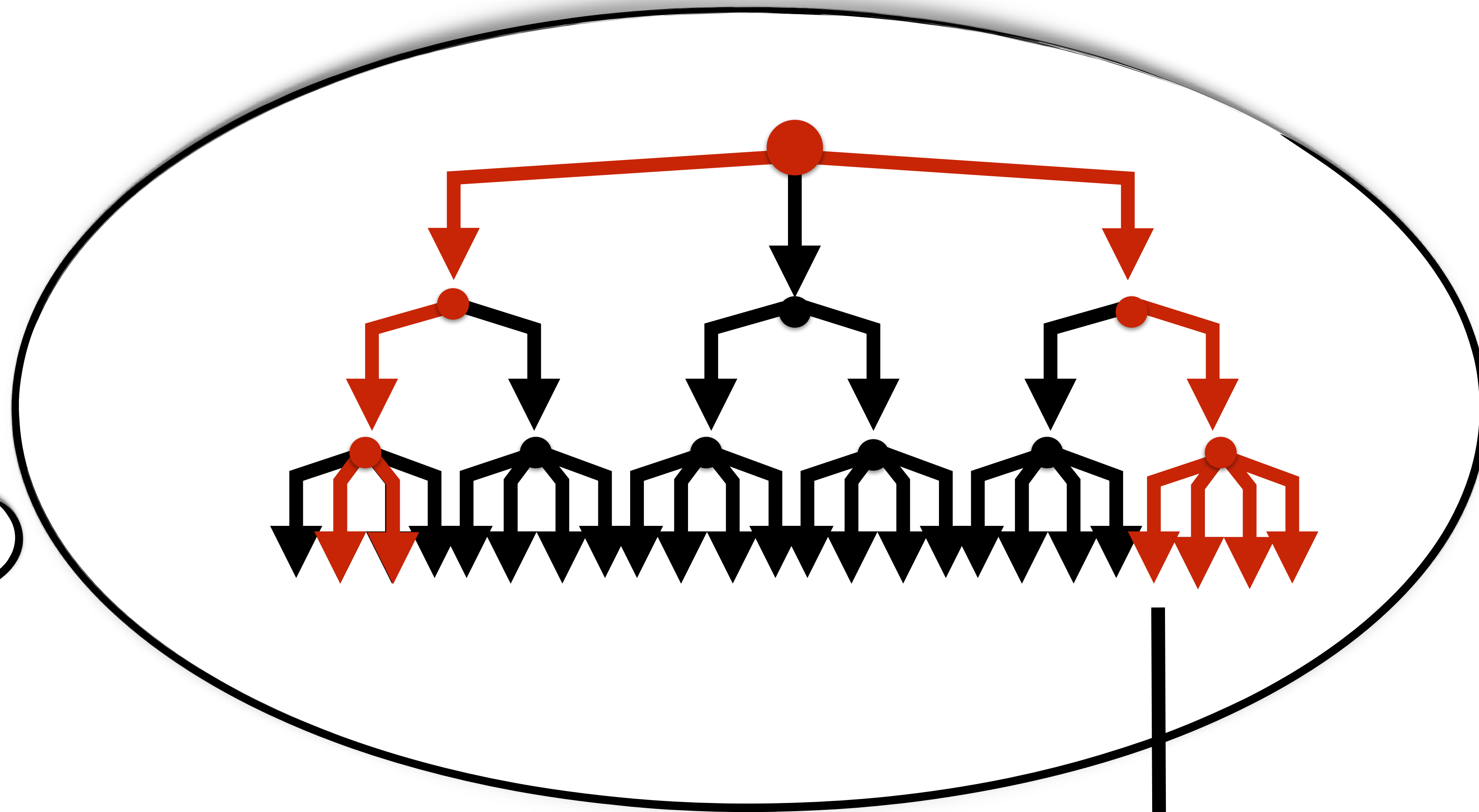
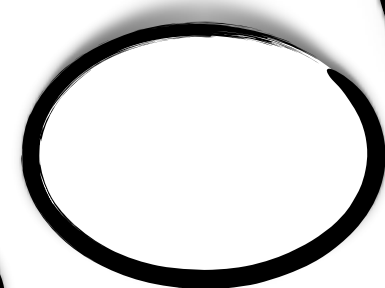
$p = 0.7$

$p = 0.07$

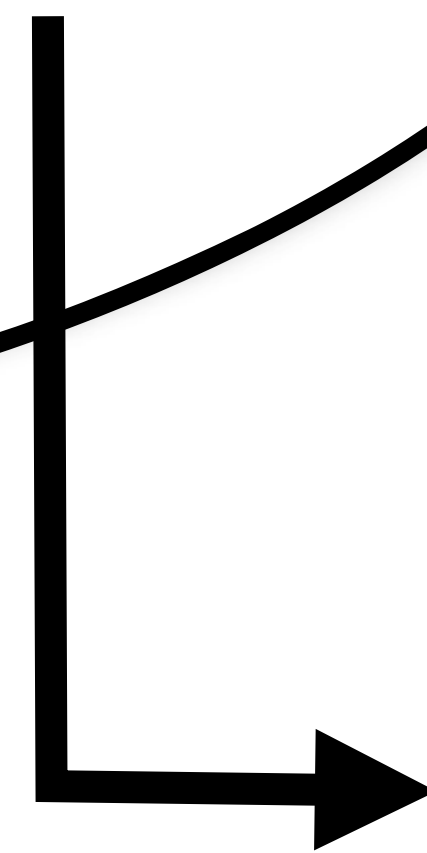








**Report all
pipelines but
do not correct**

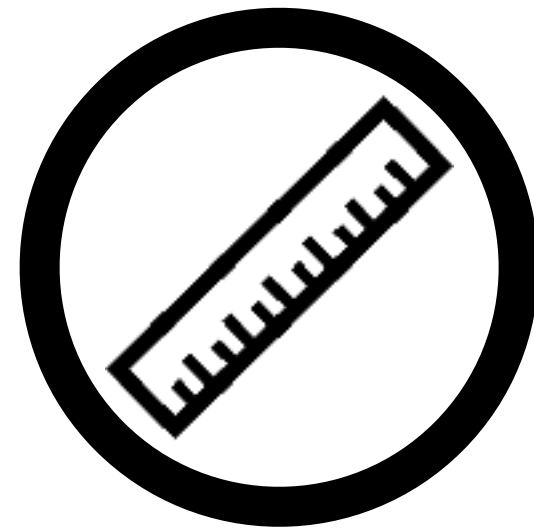


Summary





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.