Foundations of Data Science:
A/B testing

THE UNIVERSITY of EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Overview

- Principle of A/B testing
  - what it is, estimation and hypothesis testing approaches with the bootstrap

- Increasing certainty in A/B testing

- Theoretical, large-sample approach to A/B testing

- Issues in A/B testing

- Comparing numeric samples

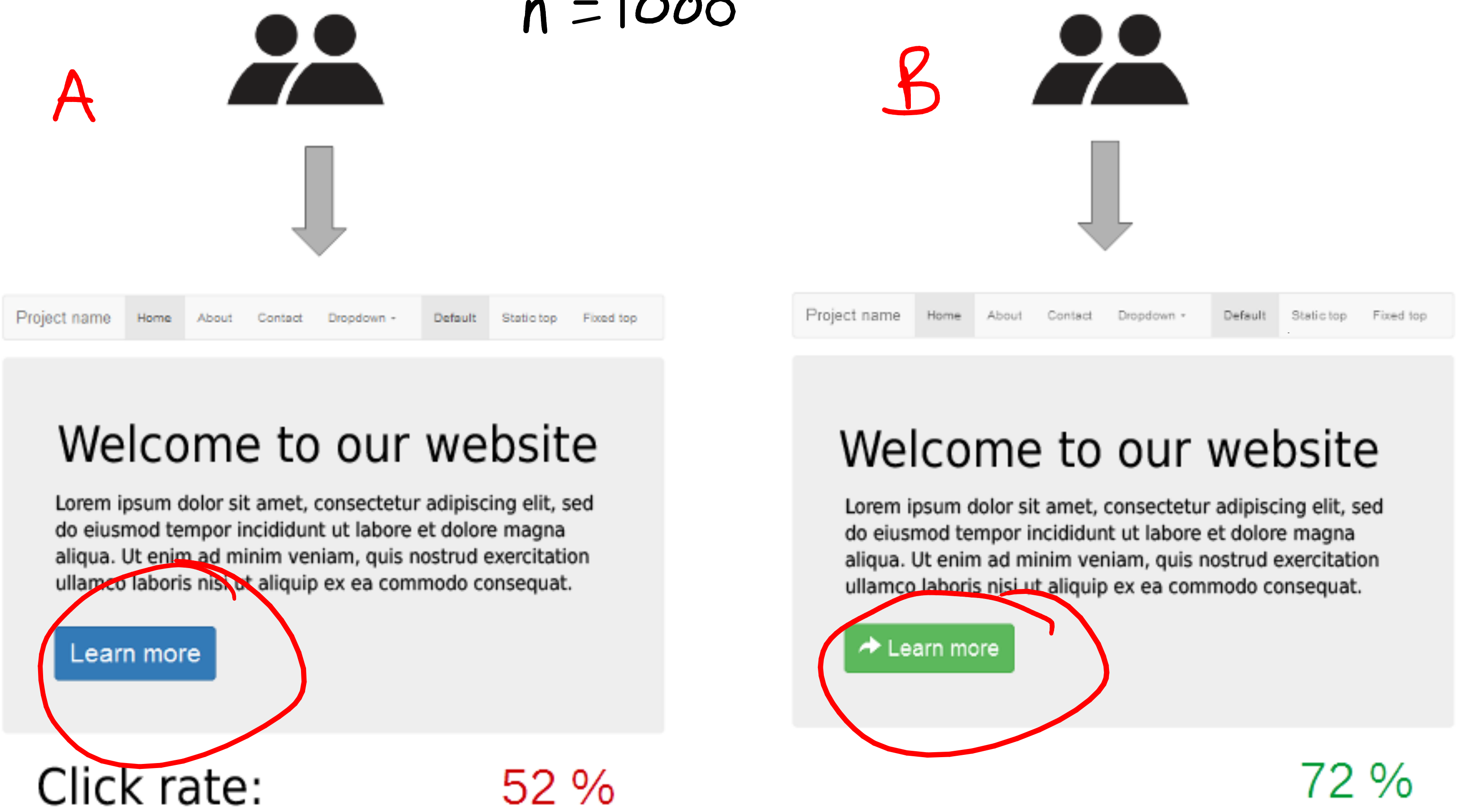
Foundations of Data Science: A/B testing - The principle of A/B testing

A/B Testing

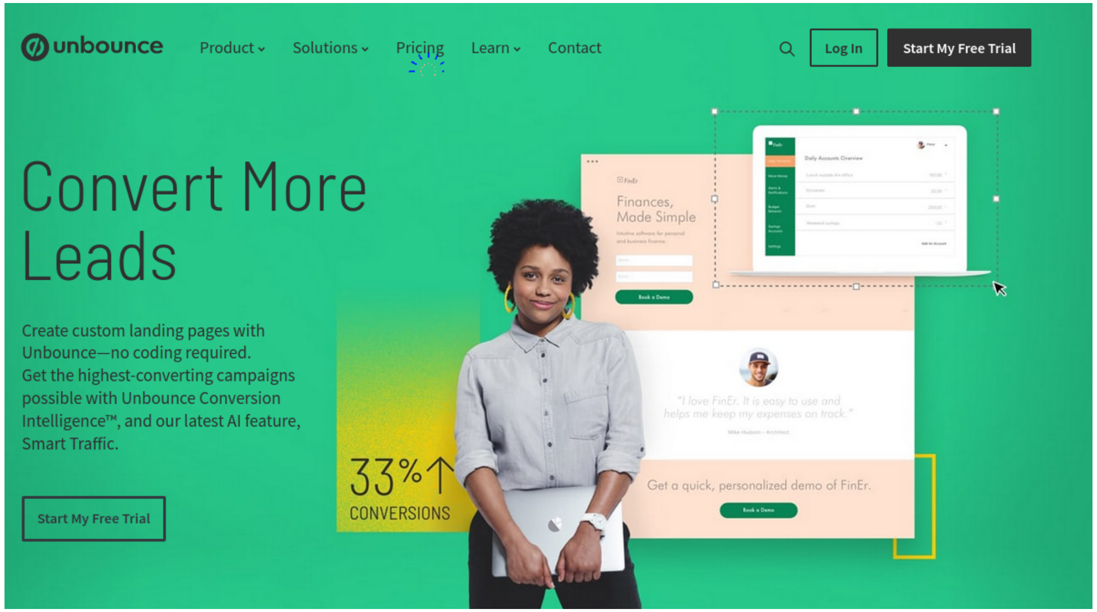
$n = 1000$

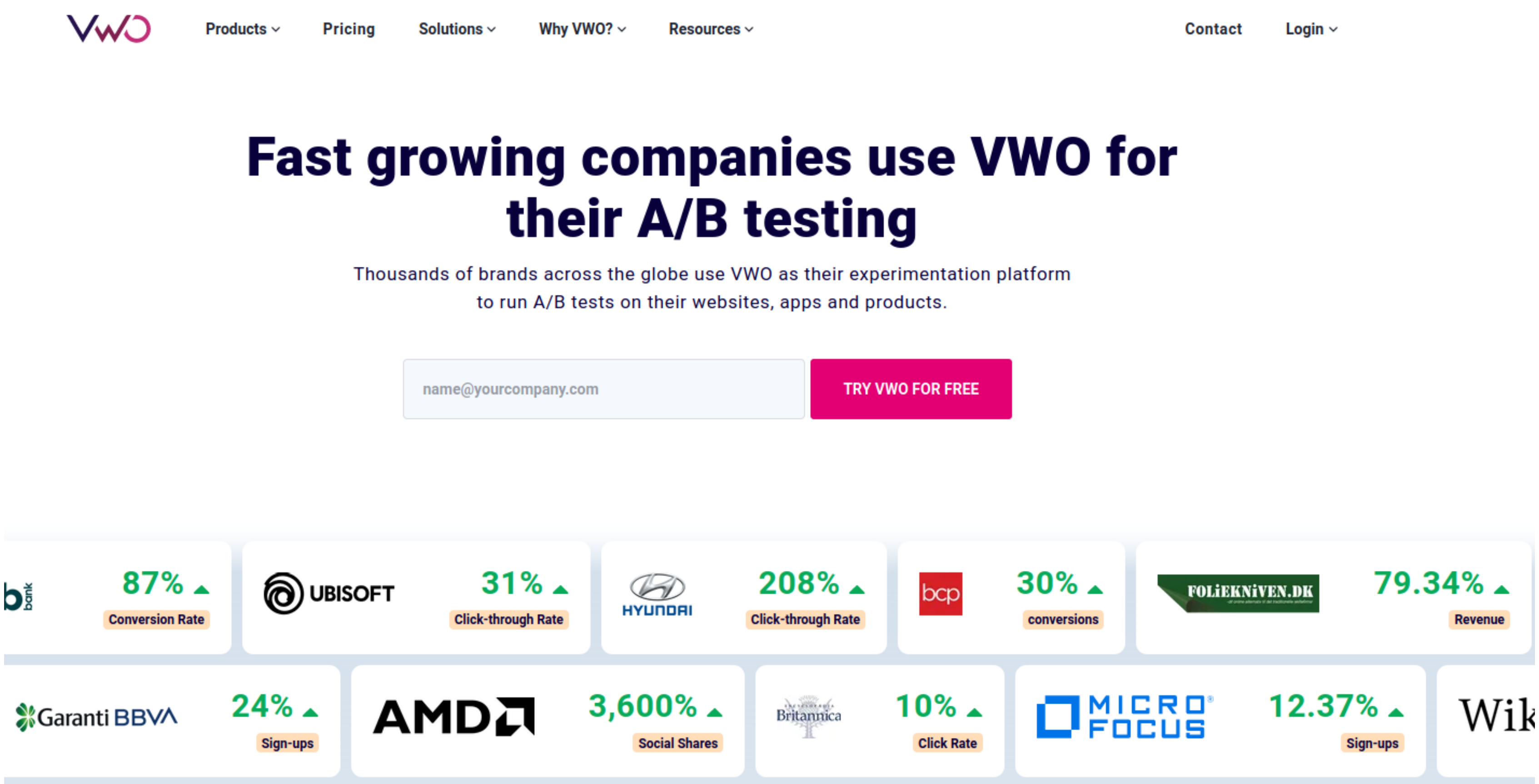
A

B



1. Is A significantly better or worse than B?
2. How much better or worse is A than B?


unbounce
Product
Solutions
Pricing
Learn
Contact
Log In
Start My Free Trial
Convert More Leads
Create custom landing pages with Unbounce—no coding required.
Get the highest-converting campaigns possible with Unbounce Conversion Intelligence™, and our latest AI feature, Smart Traffic.
Start My Free Trial
33%
CONVERSIONS
Finances, Made Simple
Get a quick, personalized demo of FinEr.


VWO
Products
Pricing
Solutions
Why VWO?
Resources
Contact
Login
Fast growing companies use VWO for their A/B testing
Thousands of brands across the globe use VWO as their experimentation platform to run A/B tests on their websites, apps and products.
name@yourcompany.com
TRY VWO FOR FREE
87%
Conversion Rate
UBISOFT
31%
Click-through Rate
HYUNDAI
208%
Click-through Rate
bcp
30%
conversions
FOLIEKNIVEN.DK
79.34%
Revenue
Garanti BBVA
24%
Sign-ups
AMD
3,600%
Social Shares
Britannica
10%
Click Rate
MICRO FOCUS
12.37%
Sign-ups

Approaches

Parameter estimation	Hypothesis testing
0. Decide underlying parameter to infer	0. Decide on H_0 and H_a
1. Construct formula for estimator in terms of data	1. Define test statistic in terms of data
2. Find approx. sampling distribution of estimator using bootstrap or large sample theory	2. Find distribution of test statistic under H_0
3. Return confidence interval	3. Reject/not reject H_0 or find p-value

A/B testing example: Estimation approach

Parameters p_A - p_B - $\}$ parameter for proportion of click-throughs from A/B

← parameter for difference.

$$d = p_A - p_B$$

Data

$n = 1000$ # presentations of A & B

$n_A = 700$ # click-throughs on A

$n_B = 720$ # " " " B

Estimators $\hat{p}_A = \frac{n_A}{n}$ $\hat{p}_B = \frac{n_B}{n}$ $\hat{d} = \hat{p}_A - \hat{p}_B$

Sampling distribution of $\hat{d}$ with bootstrap

B - # repetitions

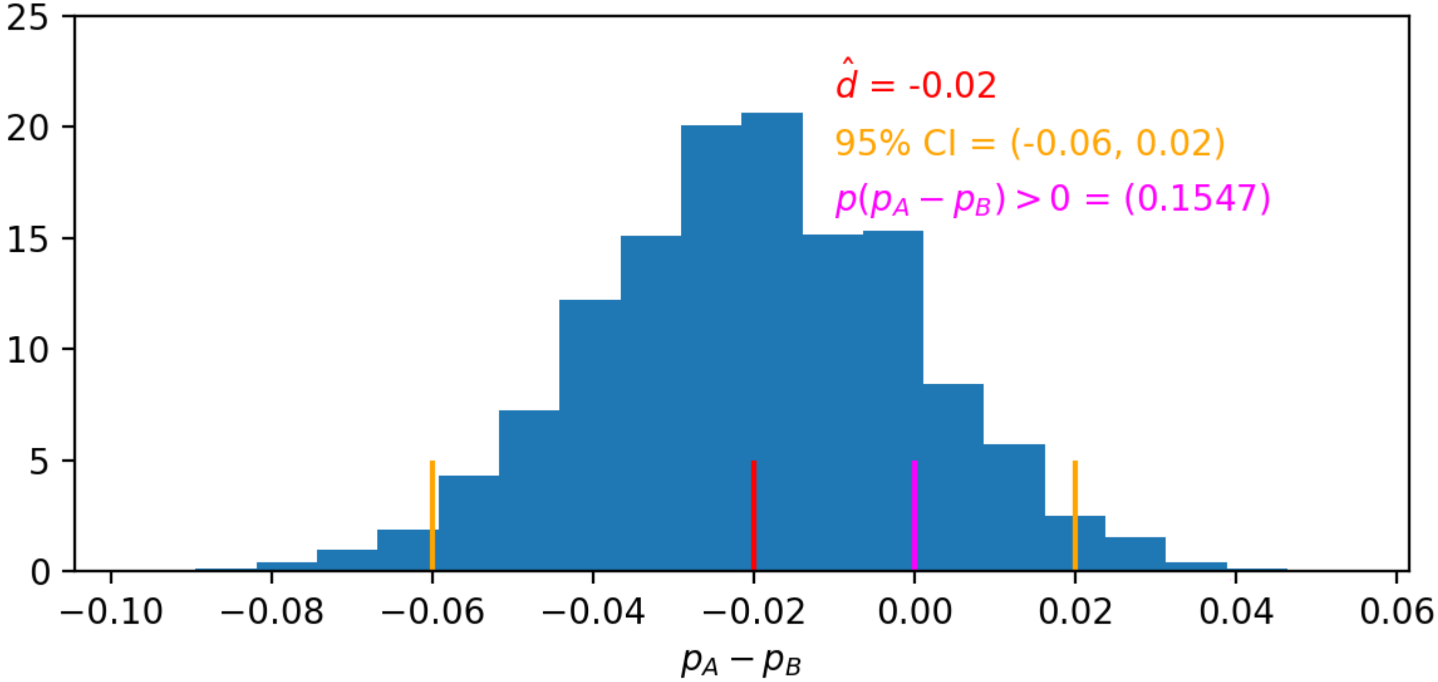
for j in 1, ..., B :

- Sample n_A^* from Binom$(n, \hat{p}_A)$
- " n_B^* " Binom$(n, \hat{p}_B)$
- Compute difference and store it.

$$d^*_j = \frac{n_A^*}{n} - \frac{n_B^*}{n}$$

Compute quantiles, std error in estimator.

Results



$$\hat{d} = \hat{p}_A - \hat{p}_B = 0.70 - 0.72 = -0.02$$

Exercise

How would you apply the hypothesis testing approach to A/B testing?

1. H_0:

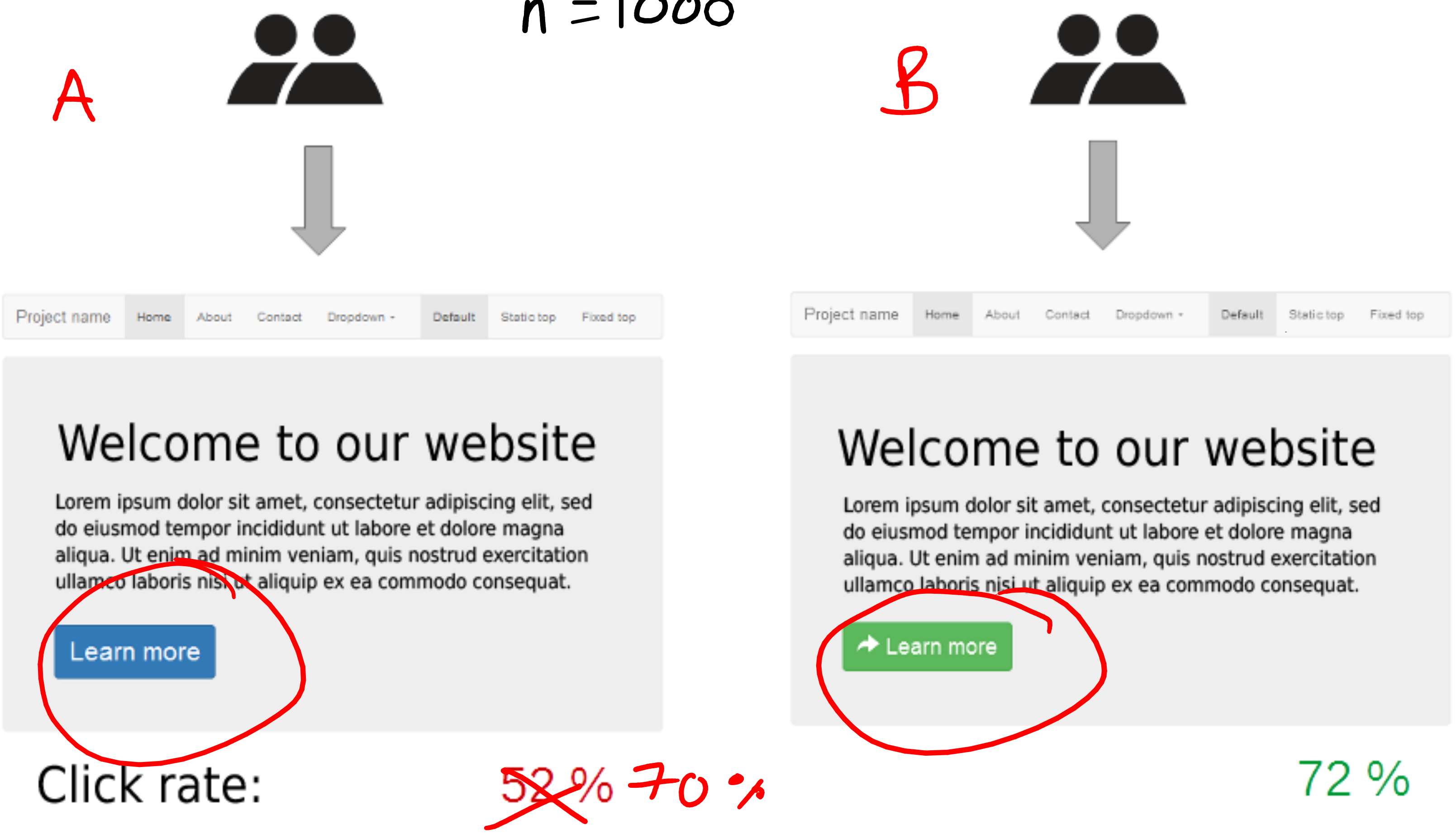
2. Test statistic:

3. Distribution of test statistic:

Foundations of Data Science: A/B testing - Increasing certainty

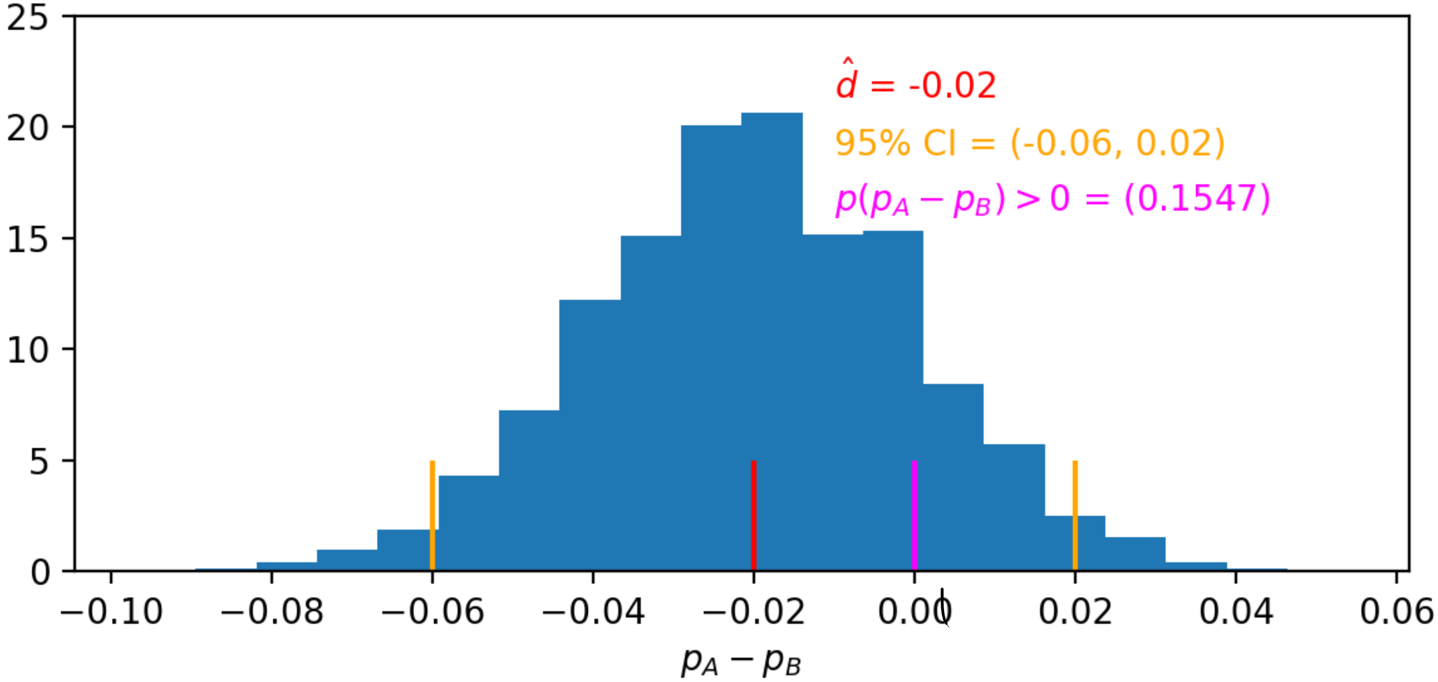
A / B Testing

n = 1000



Maxime Lorant, Wikimedia, CC SA 4.0

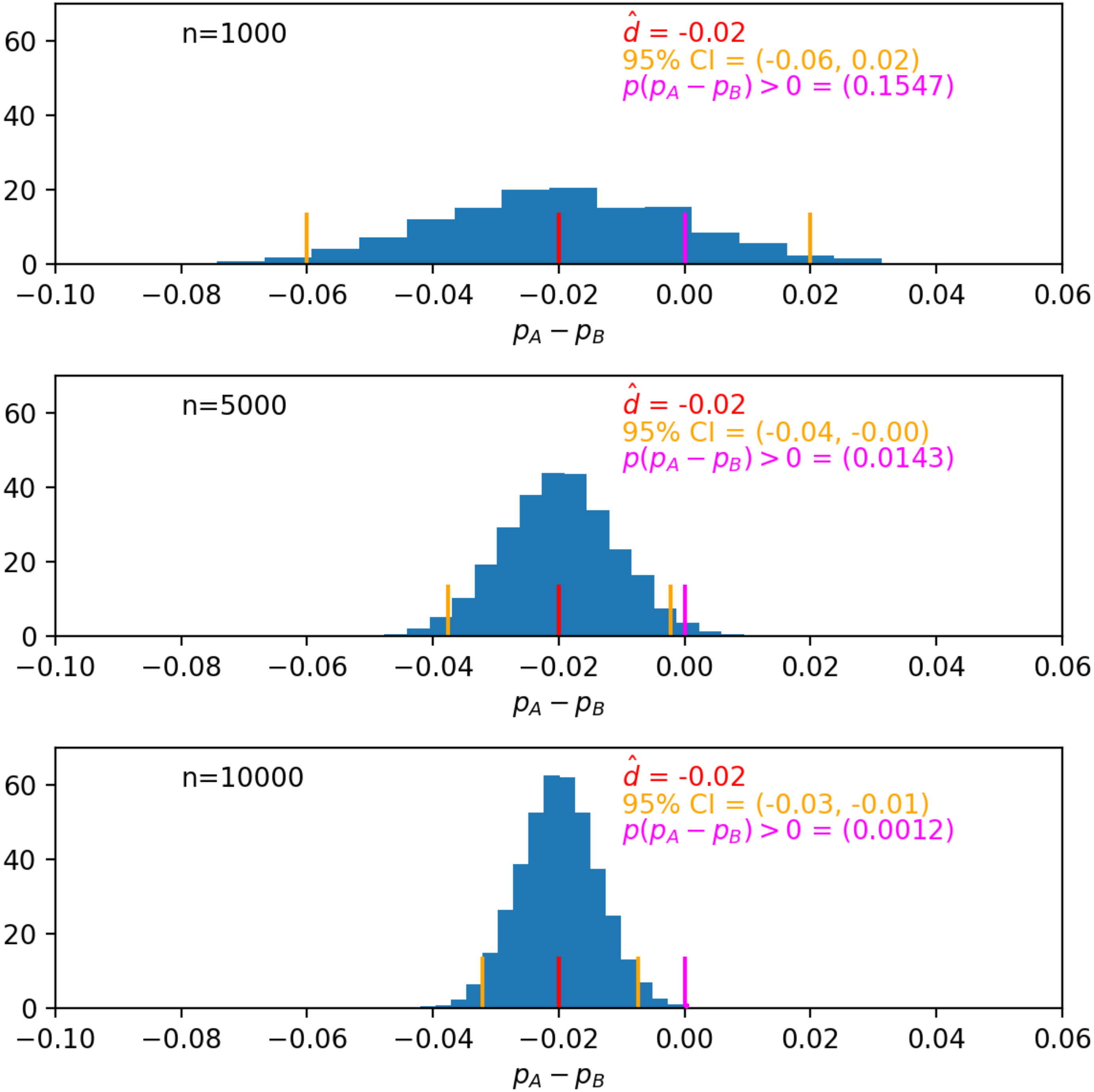
Bootstrap results



$\hat{d} = \hat{p}_A - \hat{p}_B = 0.70 - 0.72 = -0.02$

15% chance A is better than B

Getting a more certain result



Question: Is a big enough sample good enough?

We can run more experiments to get lower p-values, but could we still have the wrong answer?

Foundations of Data Science: A/B testing - Large sample theory

Confidence level : $1-\alpha$

$$CI = \left(\hat{d} - z_{\alpha/2}\,\hat{\sigma}_{\hat{d}} \,,\, \hat{d} + z_{\alpha/2}\,\hat{\sigma}_{\hat{d}}\right)$$

Eg. $\hat{d} = \hat{p}_A - \hat{p}_B = 0.70 - 0.72 = -0.02$

$$\hat{\sigma}_{\hat{d}} = \frac{\sqrt{\hat{p}_A(1-\hat{p}_A) + \hat{p}_B(1-\hat{p}_B)}}{\sqrt{n}}$$

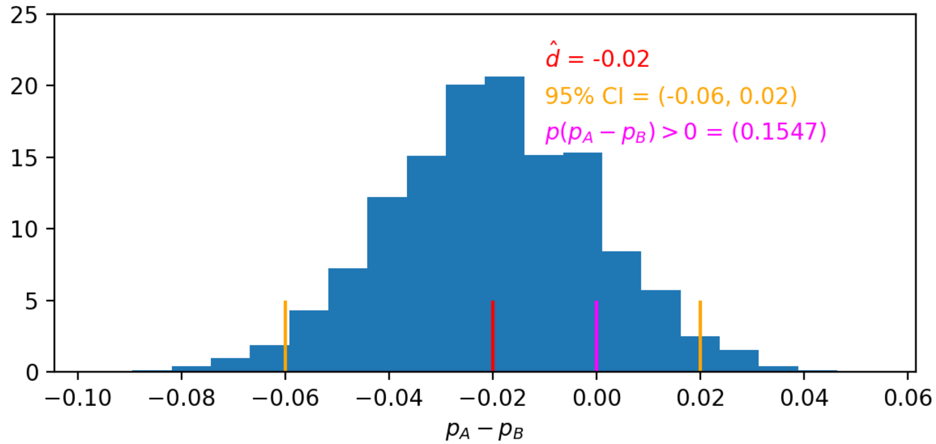
$$= \frac{\sqrt{0.70(1-0.70) + 0.72(1-0.72)}}{\sqrt{1000}} = 0.020$$

$$95\%\ CI \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96$$

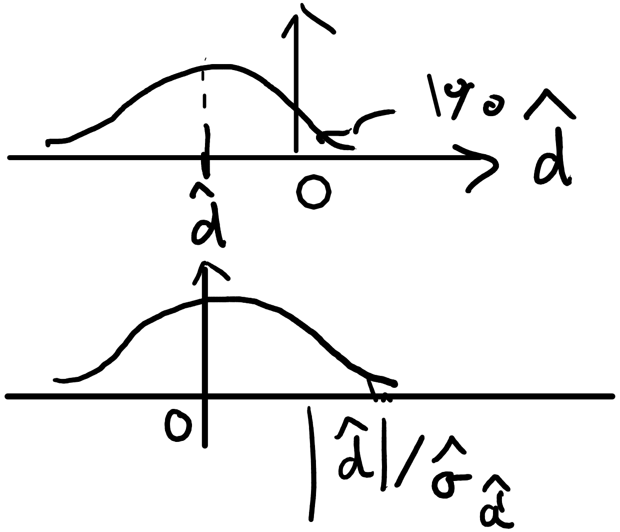
$$\Rightarrow CI : \left(\hat{d} - z_{\alpha/2}\hat{\sigma}_{\hat{d}}\ ,\ \hat{d} + z_{\alpha/2}\hat{\sigma}_{\hat{d}}\right)$$

$$= -0.02 - 1.96 \times 0.020\ ,\ 0.02 + 1.96 \times 0.02$$

$$= \underline{\underline{(-0.06\ ,\ 0.02)}}$$



Sample size calculation



$$\frac{|\hat{d}|}{\hat{\sigma}_{\hat{d}}} = z_{0.01}$$

$$\hat{\sigma}_{\hat{d}} = \frac{\sqrt{\hat{p}_A(1-\hat{p}_A) + \hat{p}_B(1-\hat{p}_B)}}{\sqrt{n}}$$

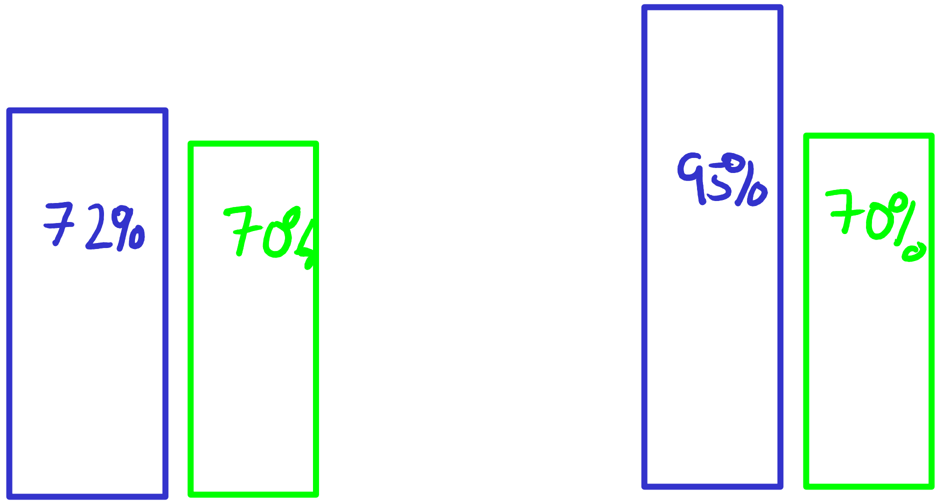
$$\frac{+\sqrt{n}\,|\hat{d}|}{\sqrt{\hat{p}_A(1-\hat{p}_A)+\hat{p}_B(1-\hat{p}_B)}} = z_{0.01}$$

$$\Rightarrow n = \frac{z_{0.01}^2}{\hat{d}^2}\left(\hat{p}_A(1-\hat{p}_A) + \hat{p}_B(1-\hat{p}_B)\right)$$

Foundations of Data Science:
A/B testing -
Issues in A/B testing

Statistical versus practical significance

Which scenario is more statistically significant?
Which scenario could be more significant practically?



$p \sim 0.001$ $\quad$ $p = 0.06$

$n = 10{,}000$ $\quad$ $n = 100$

Ethical issues

- Informed consent
  - Remember the Facebook experiment from Semester 1

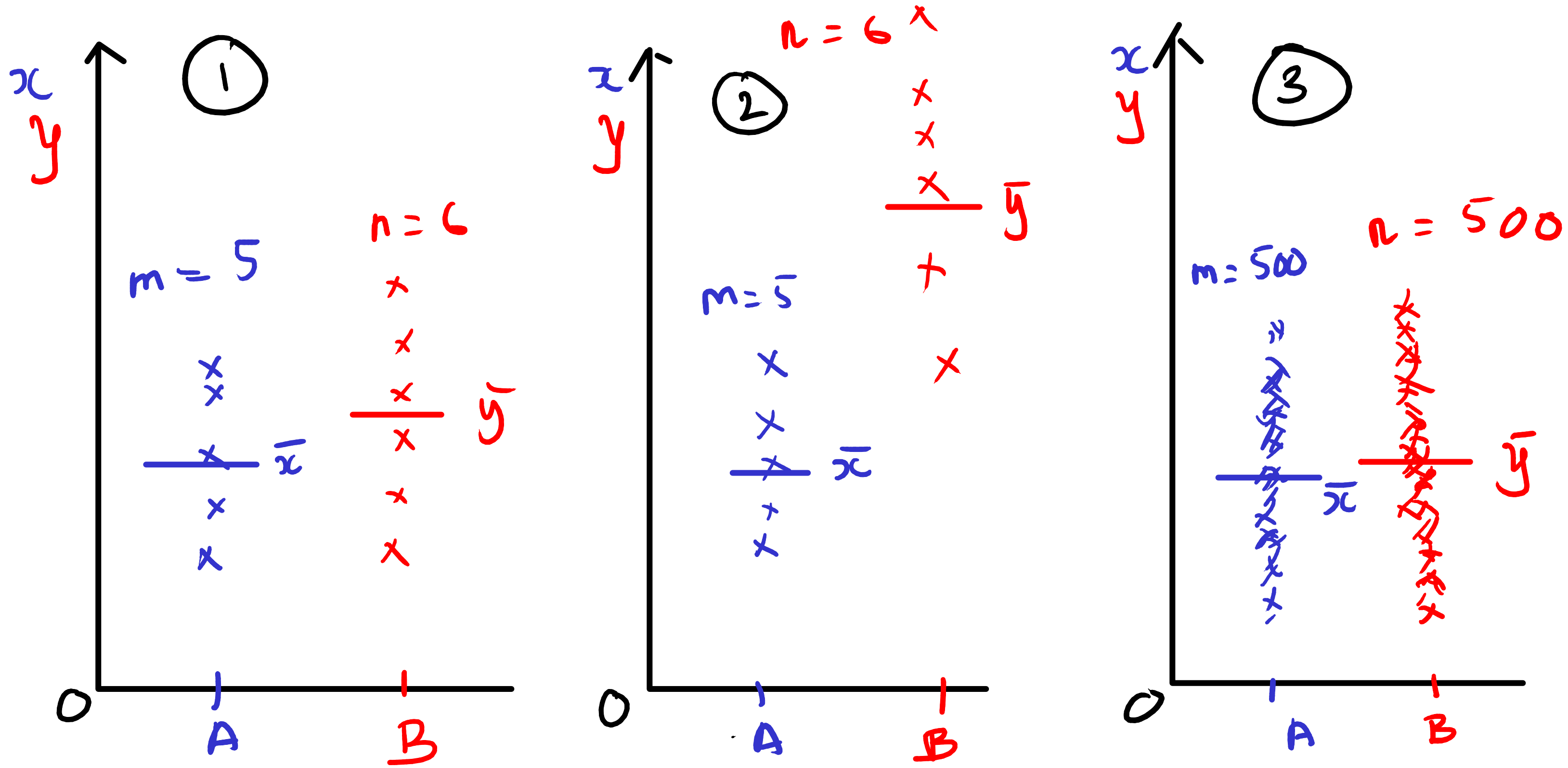
- Data protection

- Questions to ask
  - Would I feel comfortable if this change were tested on me?
  - What potential harms could be caused to users?

- Academic setting - ethics approval always needed

Foundations of Data Science: A/B testing – Comparing numeric samples

Same or different? (Hypothesis test)
How big is the difference in the means? (Estimation)



Estimator of difference: $\hat{d} = \bar{x} - \bar{y}$

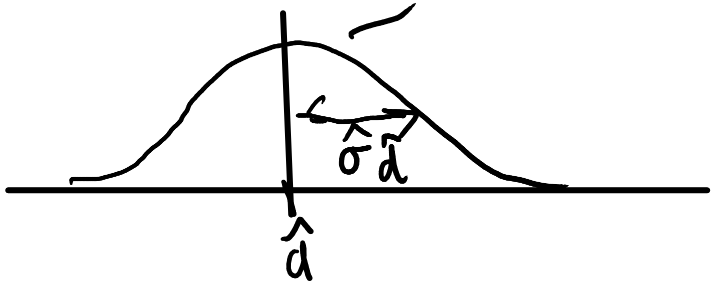
Standard error of estimator $\hat{\sigma}_{\hat{d}} = \sqrt{\dfrac{s_x^2}{m} + \dfrac{s_y^2}{n}}$

$$t = \frac{\hat{d}}{\hat{\sigma}_d}$$

Paramter estimation

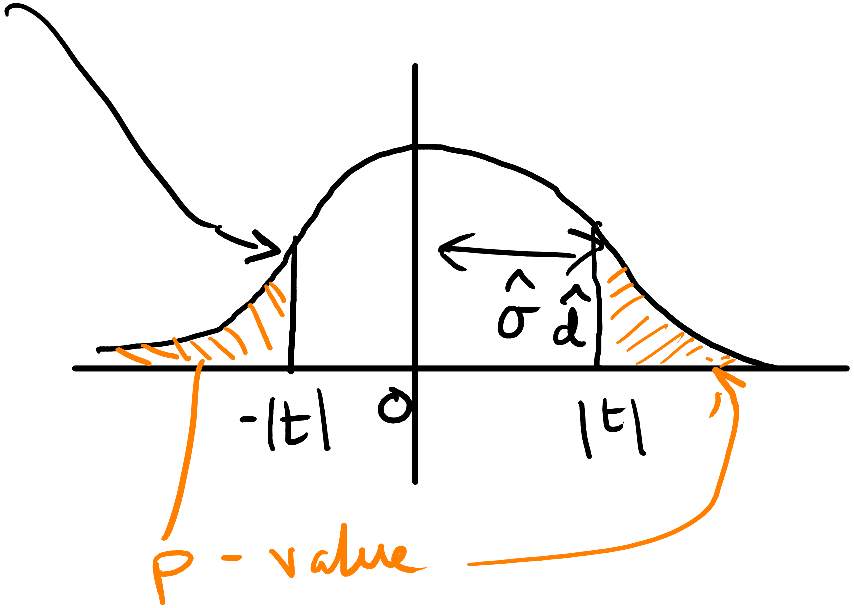
Hypothesis test (t-test)

scaled z or t-dist

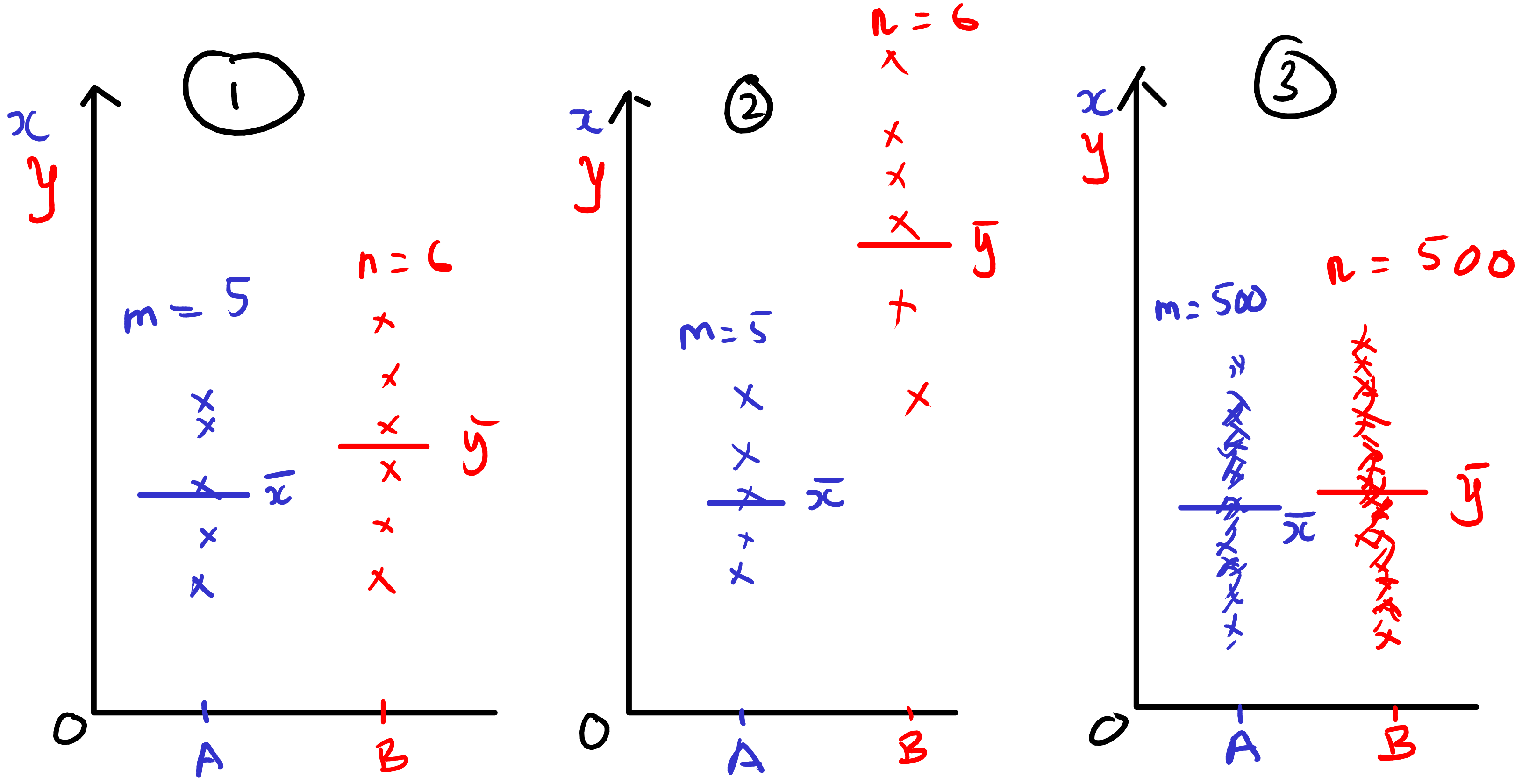


95% CI:

$$\left(\hat{d} - \hat{\sigma}_{\hat{d}}\, z_{0.025},\ \hat{d} + \hat{\sigma}_{\hat{d}}\, z_{0.025}\right)$$



Effect size - Cohen's d



$d \simeq$ - $d \simeq$ $d \simeq$

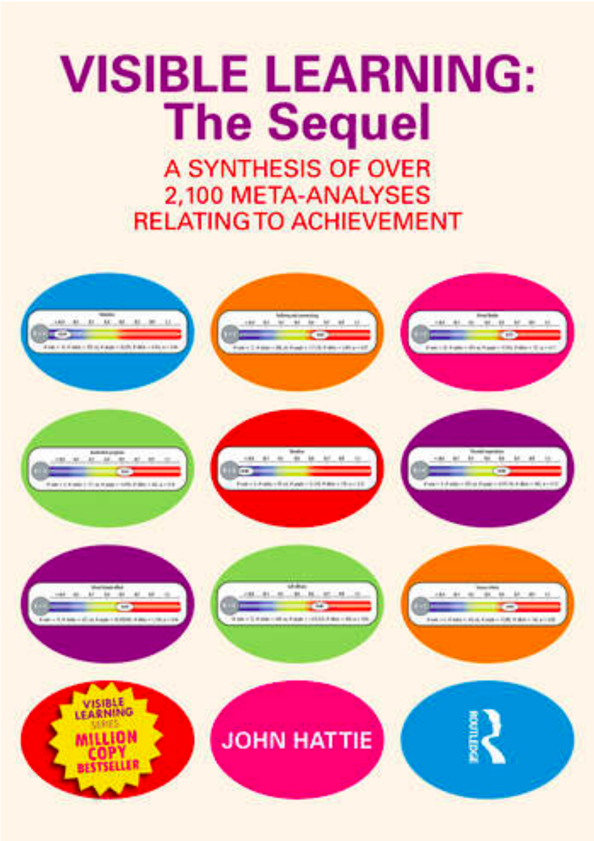
$$d = \frac{\bar{x} - \bar{y}}{s} \qquad s = \sqrt{\frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}}$$

Interpretation of Cohen's d

d=0.01	very small
d=0.2	small
d=0.5	medium
d=0.8	large
d=1.2	very large
d=2.0	huge

Cohen (1988), Sawilowsky (2009)

A well-known use of Cohen's d

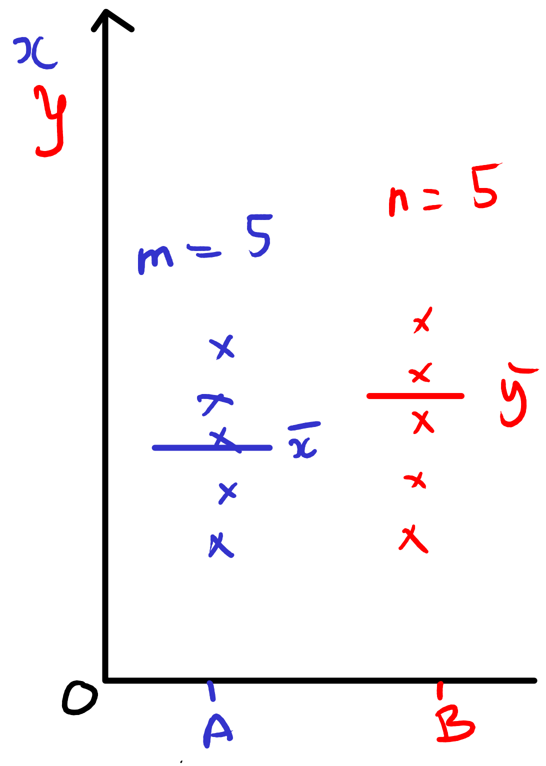


252 influences

Influence	**Cohen's *d***
Self-reported grades	1.33
Teacher credibility	0.9
Deliberate practice	0.79
Feedback	0.7
Spaced vs. mass practice	0.6
Note taking	0.5
Cooperative learning	0.4
Ability grouping for gifted students	0.3
Extra-curricula programs	0.2
Open vs. traditional classrooms	0.01
Lack of sleep	-0.05
Television	-0.18
Boredom	-0.49

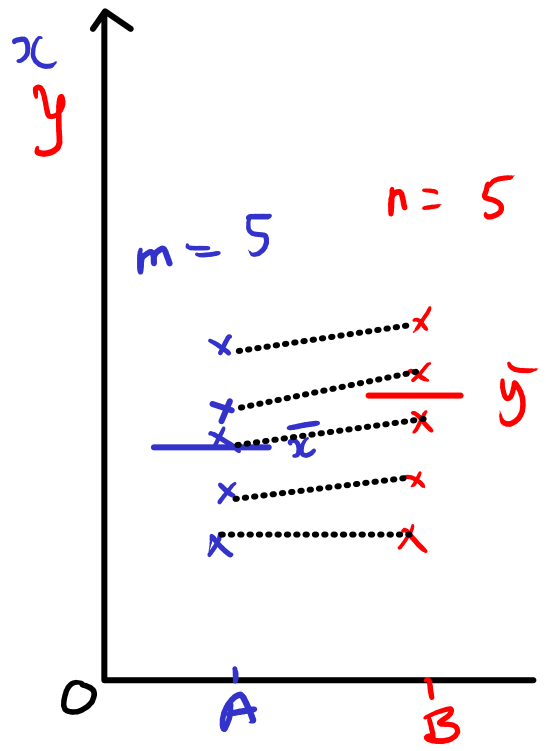
https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/

Paired data



t

p - value



t

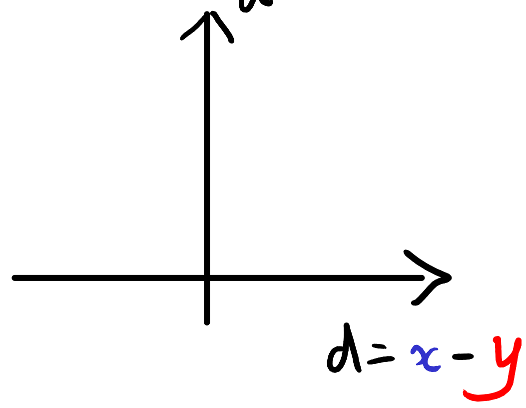
p- value

paired t–test

$$d_i = x_i - y_i$$

$$\hat{\sigma}_{\hat{d}} = \sqrt{\frac{1}{n} \sum_i (x_i - y_i)^2}$$

$$t = \frac{\bar{d}}{\hat{\sigma}_{\hat{d}}}$$



Summary

1. A/B testing: controlled experiment, binary response

2a. Estimate confidence intervals between response rates in A and B, by bootstrap or theoretically

b. Test if response rate in A is different from B, by statistical simulation, or theoretically

3. Increasing sample size decreases confidence interval and and decreases p-value

4. Issues: Ethics and effect size

5. Numeric samples - estimation, hypothesis testing, effect size (Cohen's d)