Inf2 - Foundations of Data Science: A/B testing

THE UNIVERSITY of EDINBURGH informatics

FOUNDATIONS OF DATA SCIENCE

Announcement

No comprehension questions for this lecture.... yet

But remember Week 4 and Week 6 sorkshops on statistical problems

Plan for statistical inference

1. Randomness, sampling and simulations (S2 Week 1)

2. Estimation, including confidence intervals (S2 Week 2)

3. Hypothesis testing (S2 Week 3)

4. A/B testing (S2 Week 3)

Onwards to Logistic regression (S2 Week 4)

Today

- 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

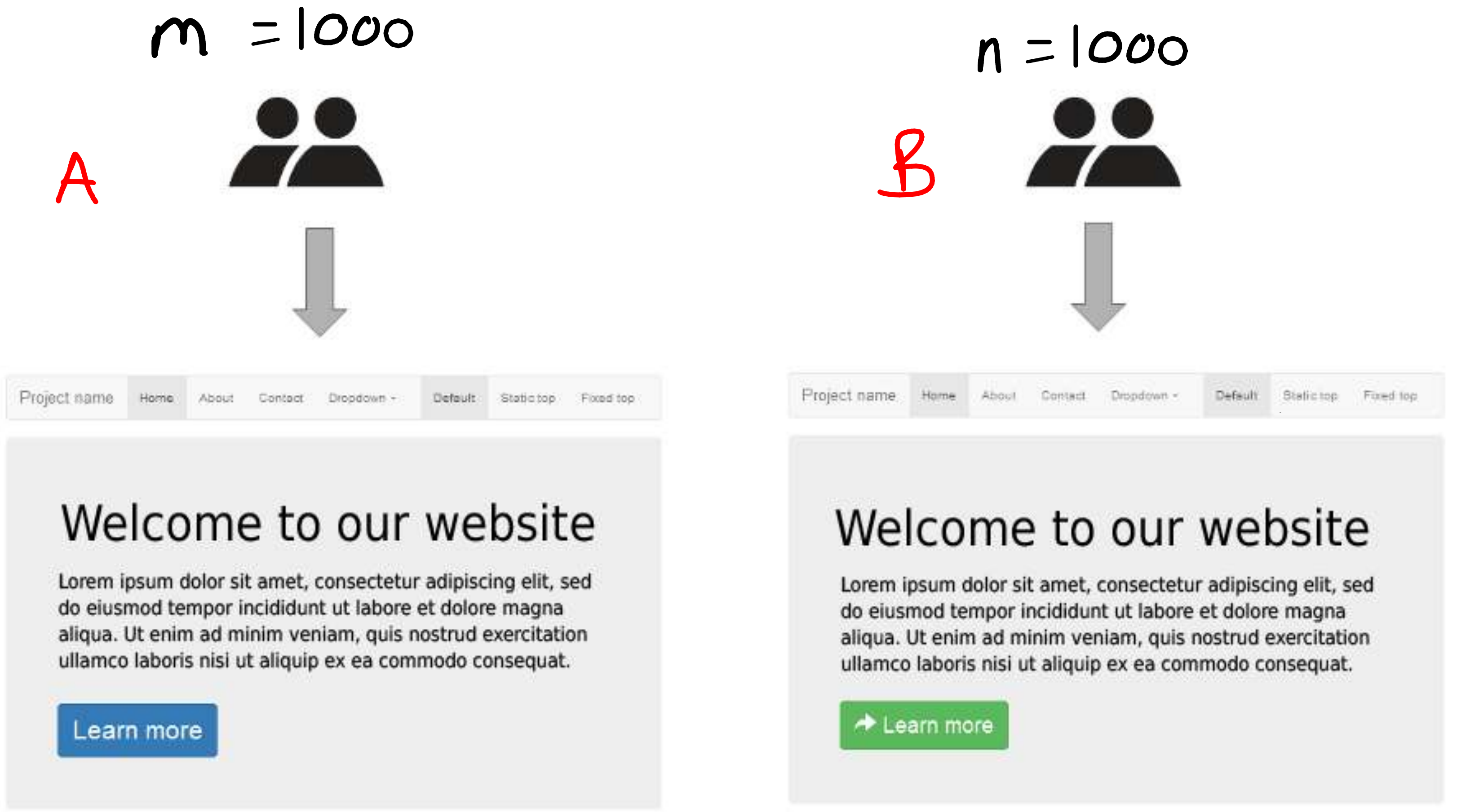
- Comparing paired samples

Inf2 - Foundations of Data Science: The principle of A/B testing

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

A/B Testing = Randomised controlled trial

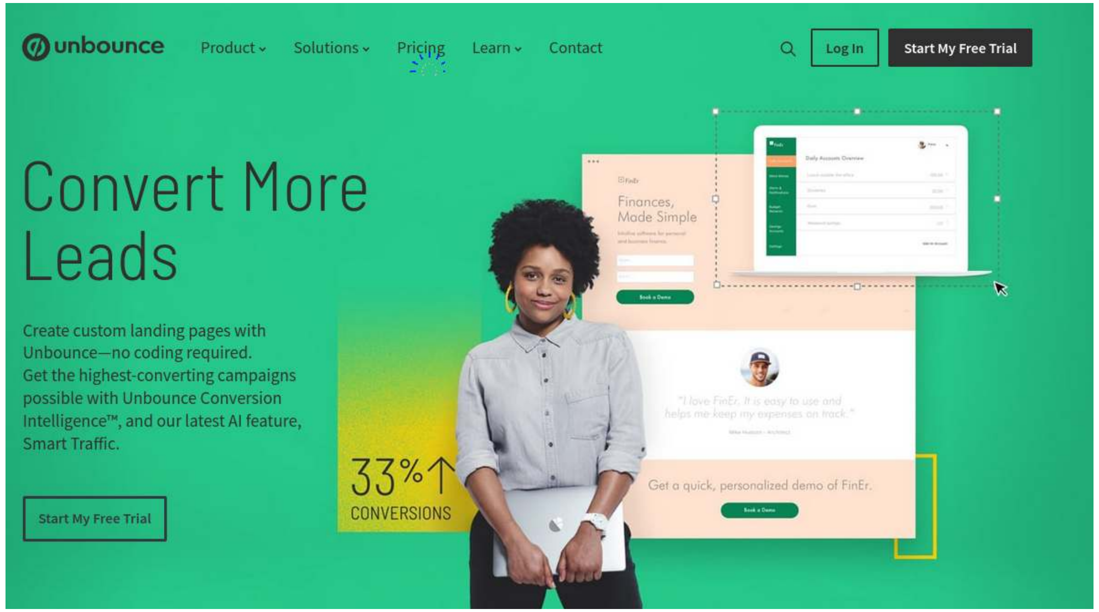


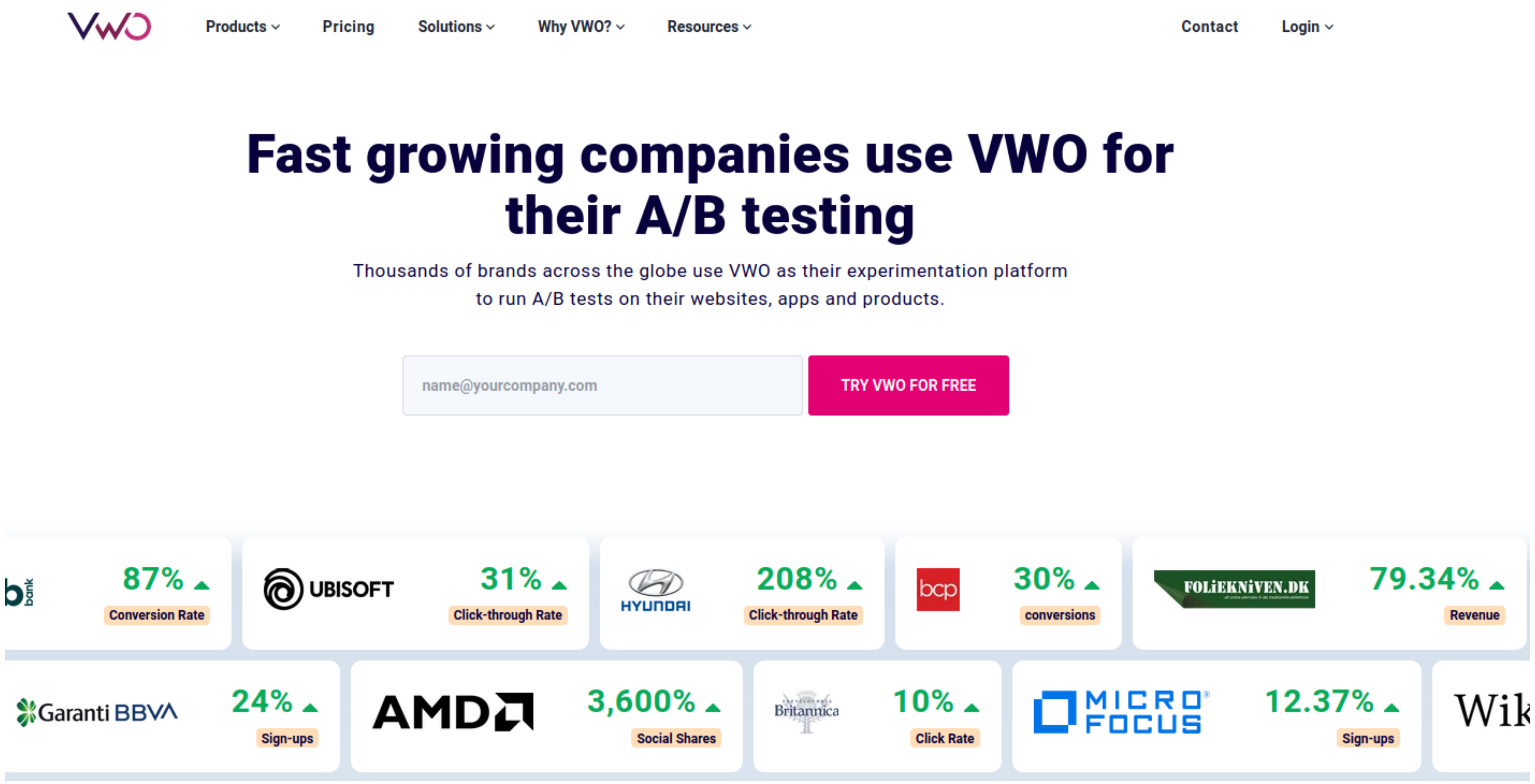
Treatment

Click rate: 52 % 72 %

Response

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

A/B testing example: Estimation approach

Parameters

- } parameter for proportion of click-throughs from A/B
- ← parameter for difference.

Data

$m = 1000$ # presentations of A

$n = 1000$ # presentations of B

$= 700$ # click-throughs on A

$= 720$ # " " " B

Estimators

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

B - # repetitions

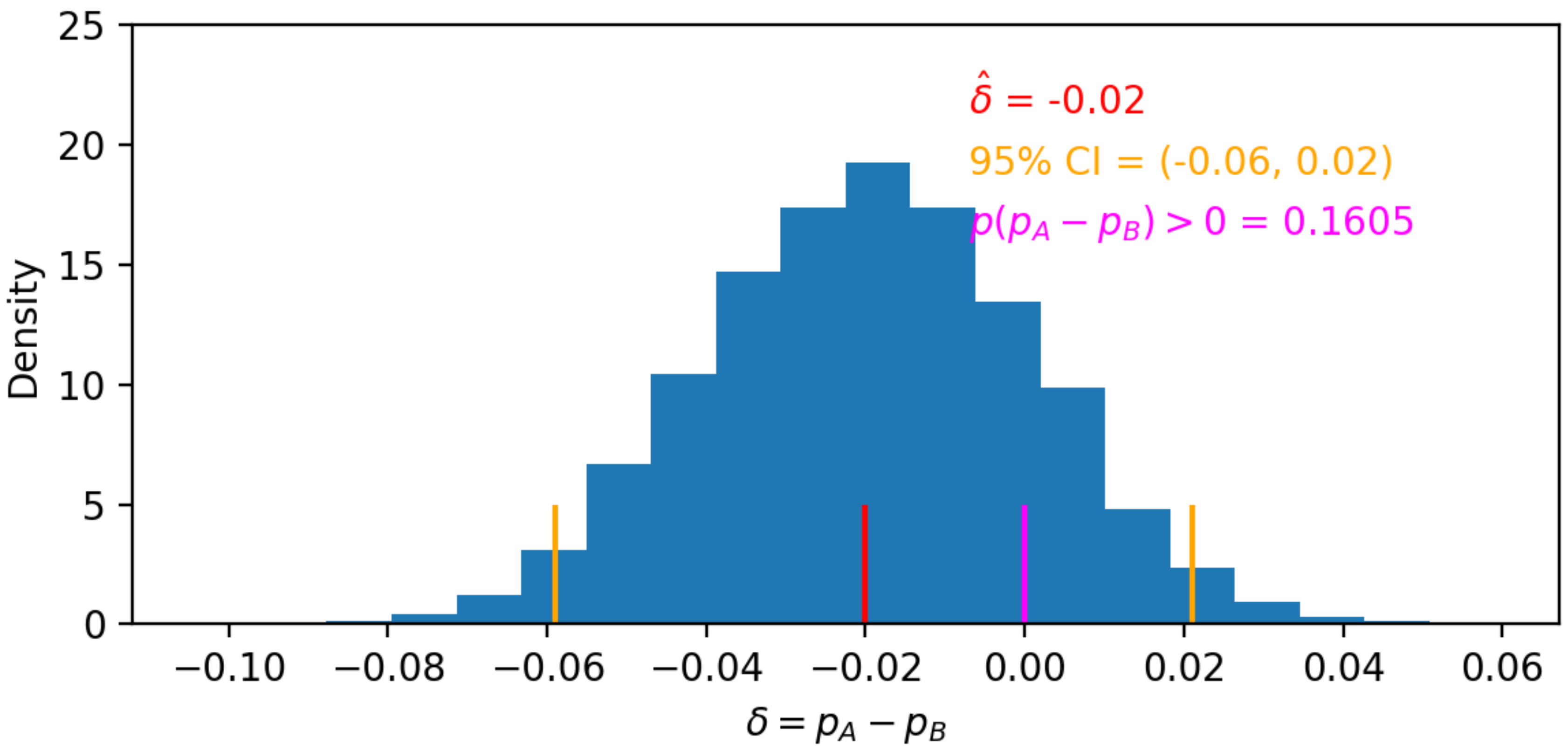
for j in 1, ..., B :

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

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

Compute quantiles, std error in estimator.

Results



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

Hypothesis testing approach

H_0 : Group A and Group B have same click-through rate $p = \frac{n_A + n_B}{m + n}$

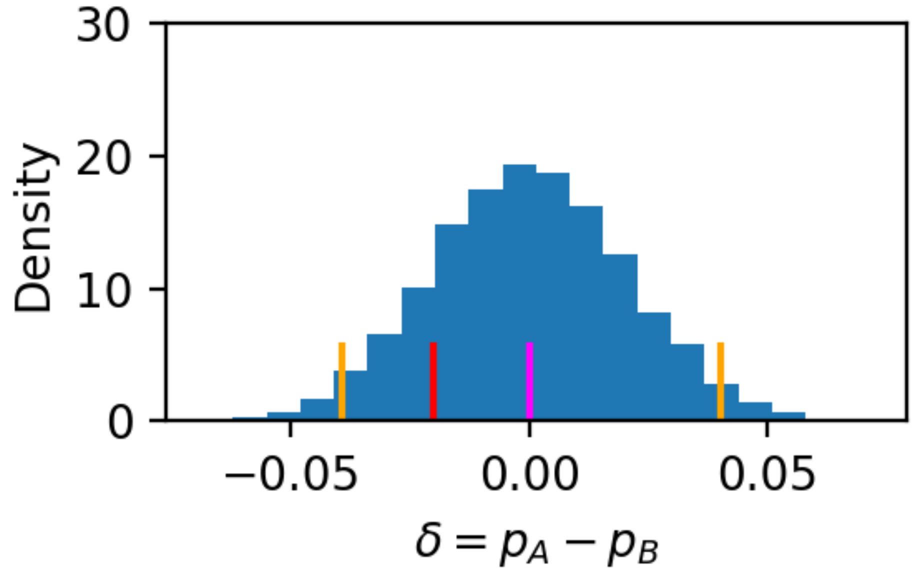
Test statistic : difference $\Delta = \frac{N_A}{m} + \frac{N_B}{n}$

Statistical model : $N_A \sim \text{Binom}(m, p)$

$N_B \sim \text{Binom}(n, p)$

$$\Delta = \frac{N_A}{m} - \frac{N_B}{n}$$

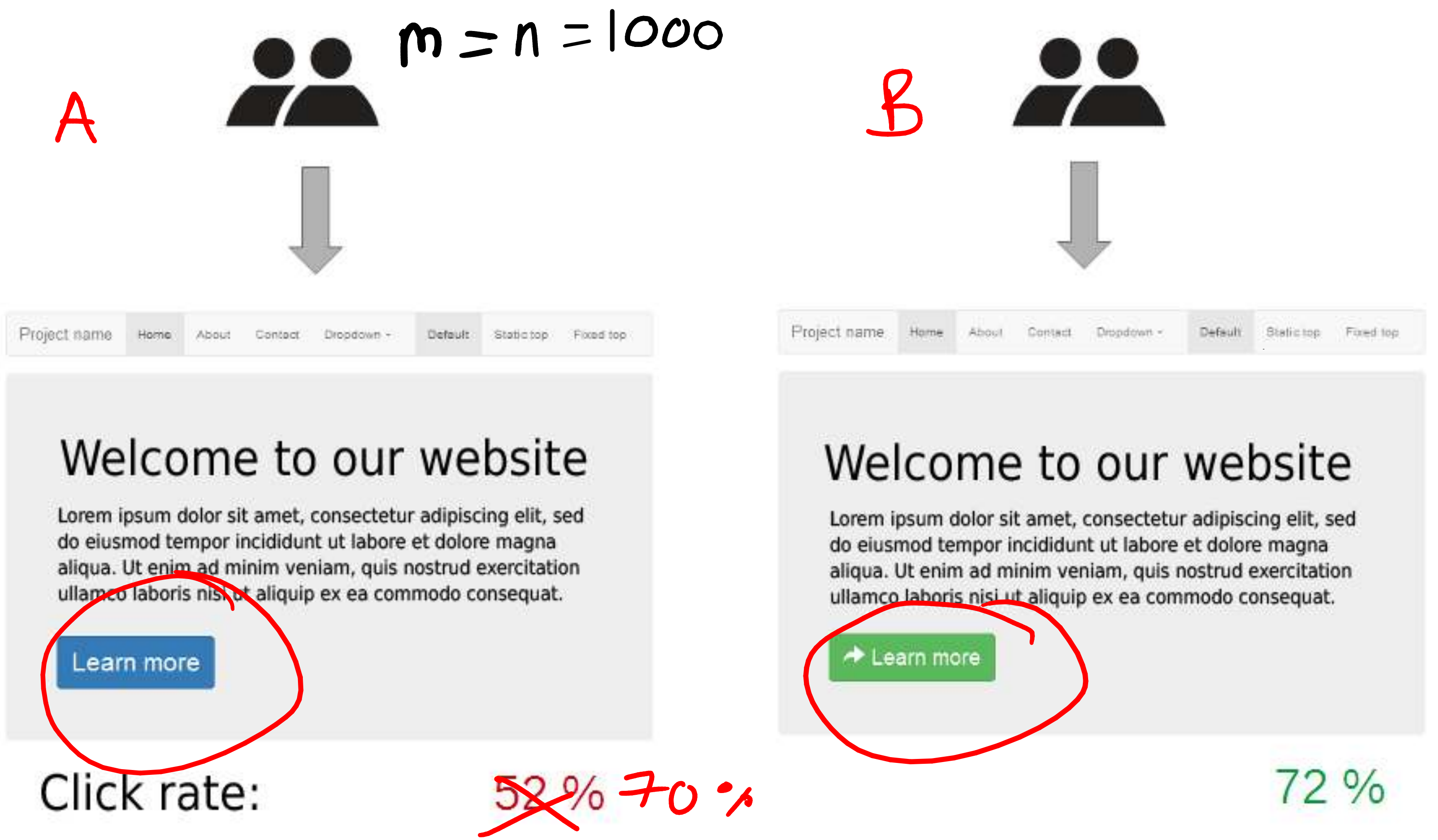
Compare with $\delta = \frac{n_A}{m} - \frac{n_B}{n}$



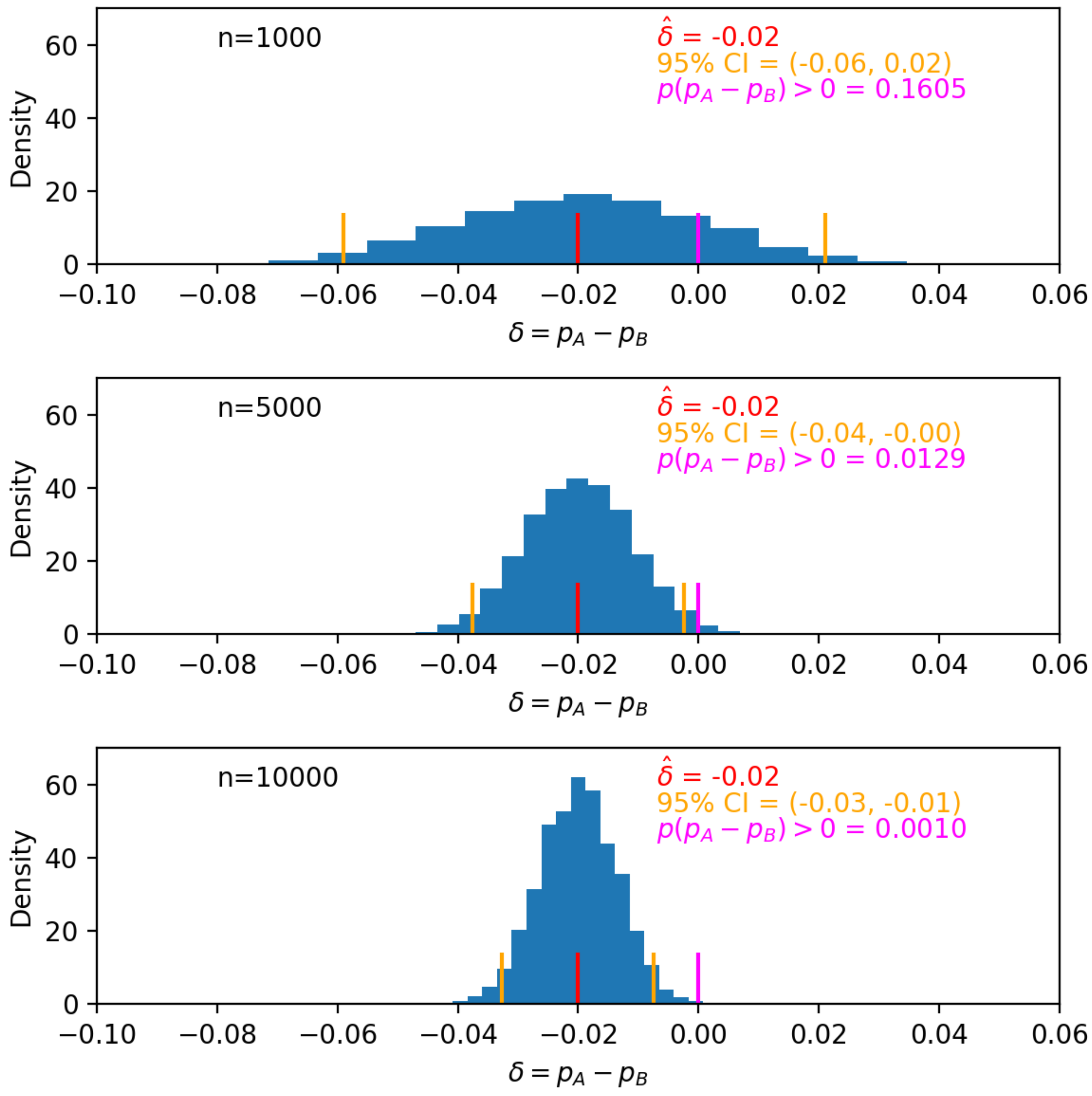
Inf2 - Foundations of Data Science:
A/B testing -
Increasing certainty

THE UNIVERSITY of EDINBURGH informatics

FOUNDATIONS OF DATA SCIENCE


m = n = 1000
A
B
Project name
Home
About
Contact
Dropdown
Default
Static top
Fixed top
Welcome to our website
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.
Learn more
Click rate:
52 %
70 %
72 %

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?

Inf2 - Foundations of Data Science:
A/B testing -
Comparing groups with numeric responses

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Notation and example

Sample A : $x_1, \ldots, x_m$
Sample B : $y_1, \ldots, y_n$

Estimation

Estimation: CI for $\delta = \mu_x - \mu_y$
Estimator: $\hat{\delta} = \bar{X} - \bar{Y}$

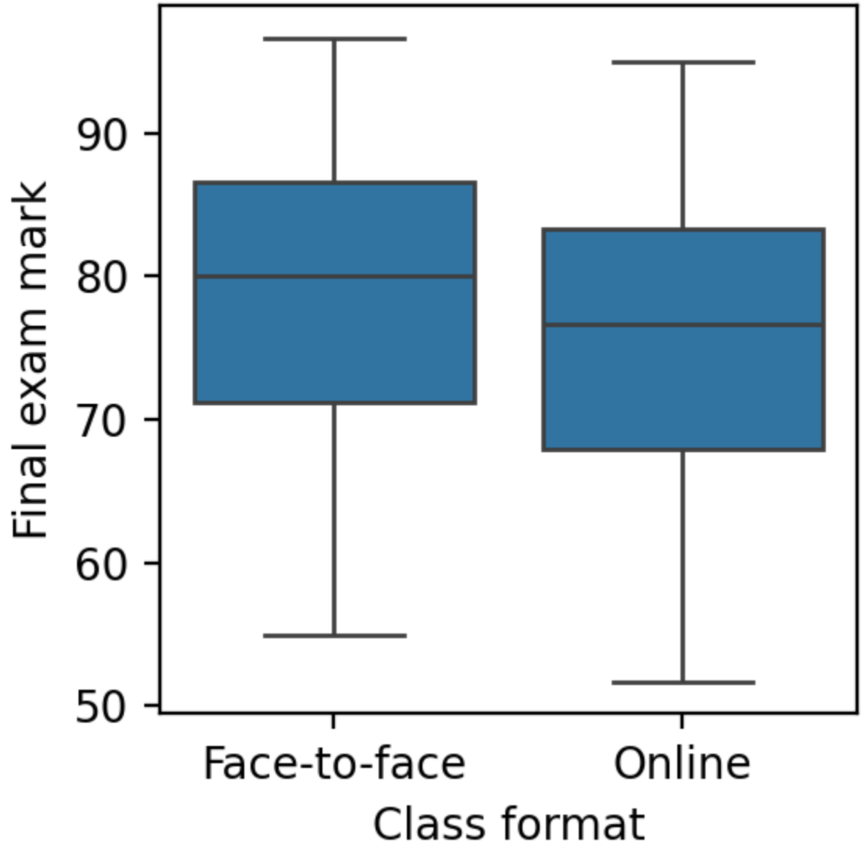
Hypothesis testing: .

H_0 : $\mu_x - \mu_y = 0$

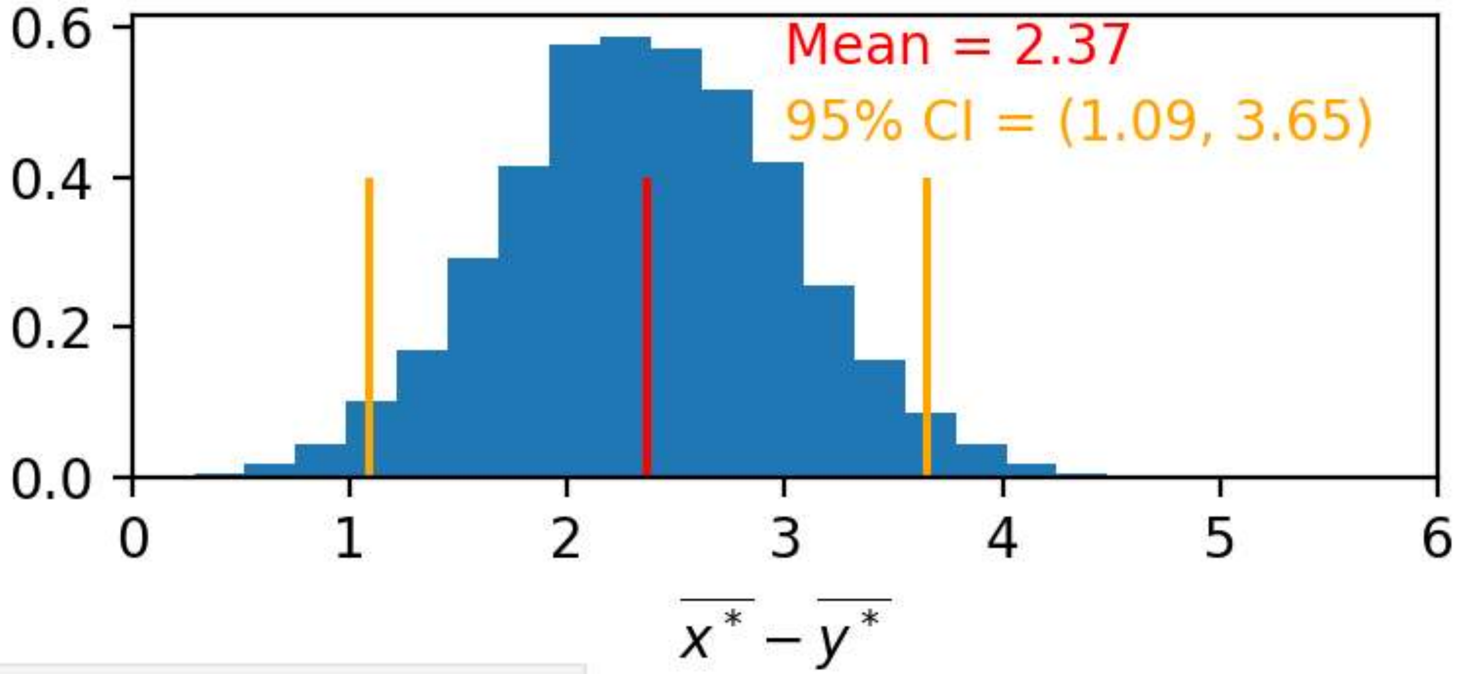
under various assumptions about variances.

Alpert et al. (2016) RCT of face-to-face versus online learning

$m = 80$ $n = 74$



Bootstrap simulation




```
B = 10000
m = len(x)
n = len(y)
dstar = pd.Series(B, dtype=float)
for j in range(B):
    xstar = x.sample(n, replace=True)
    ystar = y.sample(n, replace=True)
    dstar.loc[j] = xstar.mean() - ystar.mean()

ci95 = dstar.quantile([0.025, 0.975])
```

Theoretical method

Estimator $\hat{\delta} = \bar{X} - \bar{Y}$

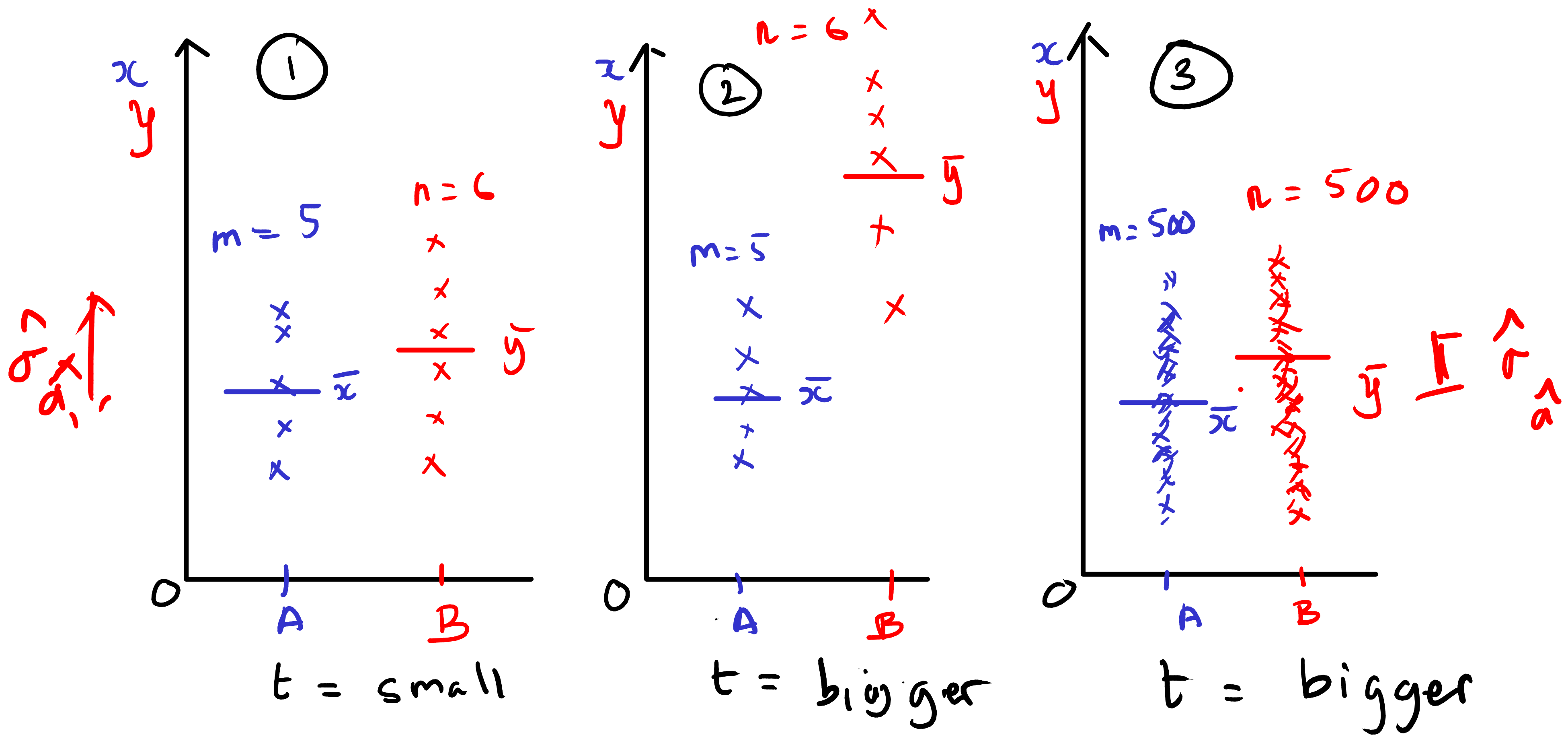
$Var(\bar{X}) = \dfrac{\sigma_x^2}{m}$ $\quad$ $Var(\bar{Y}) = \dfrac{\sigma_y^2}{n}$

$\sigma^2_{\hat{\delta}} = Var(\bar{X} - \bar{Y}) = Var(\bar{X}) - Var(\bar{Y}) = \dfrac{\sigma^2_x}{m} + \dfrac{\sigma^2_y}{n}$

Estimator $\hat{\sigma}^2_{\hat{\delta}} = \dfrac{S_x^2}{m} + \dfrac{S_y^2}{n}$

Expect $T = \dfrac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_x^2/m + S_y^2/n}}$ to have z-dist for large n, m > 40

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

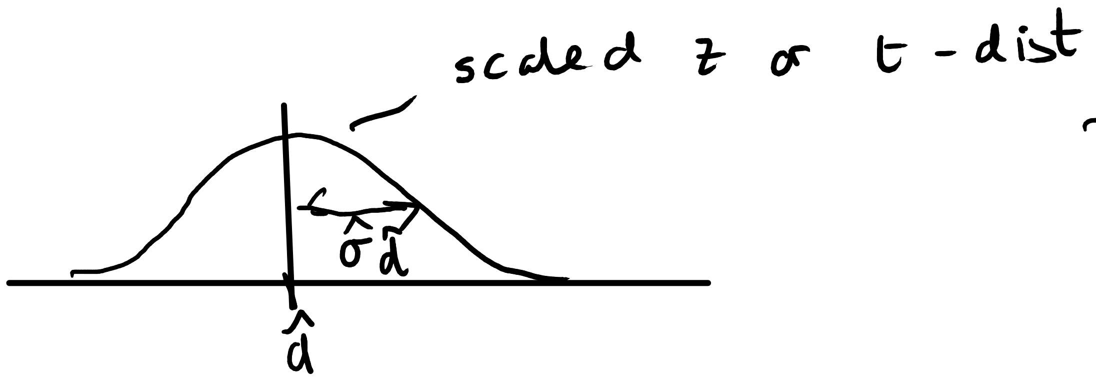


Estimator of difference: $\hat{\delta} = \bar{x} - \bar{y}$

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

$t = \frac{\hat{\delta}}{\hat{\sigma}_{\hat{\delta}}}$

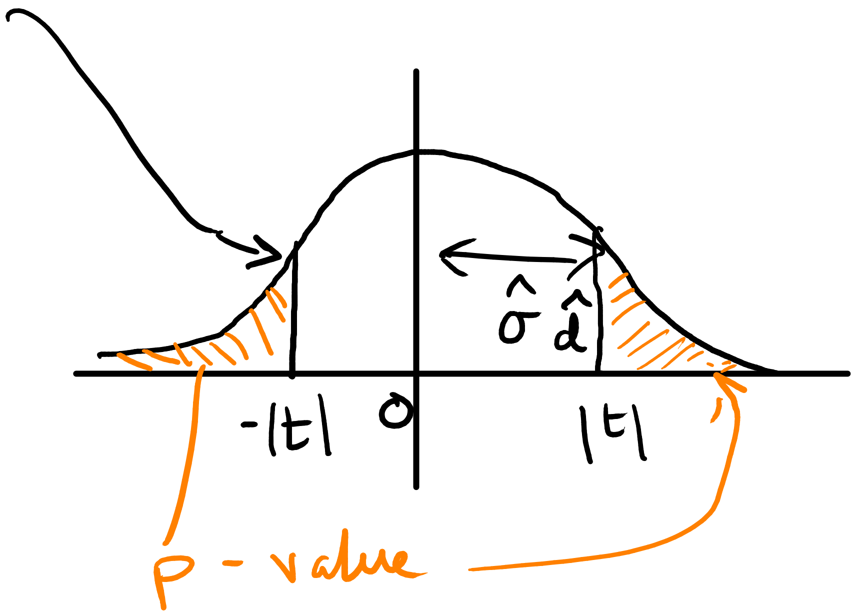
Parameter estimation



95% CI :

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

Hypothesis test (t-test)



Proportions

$$x_1, \dots x_m = 0, 1, 1, \dots, 0, 1, 1, 1, 0$$

$$\bar{x} = \hat{p}_A$$

$$s_x^2 \approx \hat{p}_A(1-\hat{p}_A)$$

$$\bar{y} = \hat{p}_B$$

$$s_y^2 = \hat{p}_B(1-\hat{p}_B)$$

$$\Rightarrow \hat{\sigma}_{\bar{x}-\bar{y}} = \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{m} + \frac{\hat{p}_B(1-\hat{p}_B)}{n}}$$

Worked example of proportions

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

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

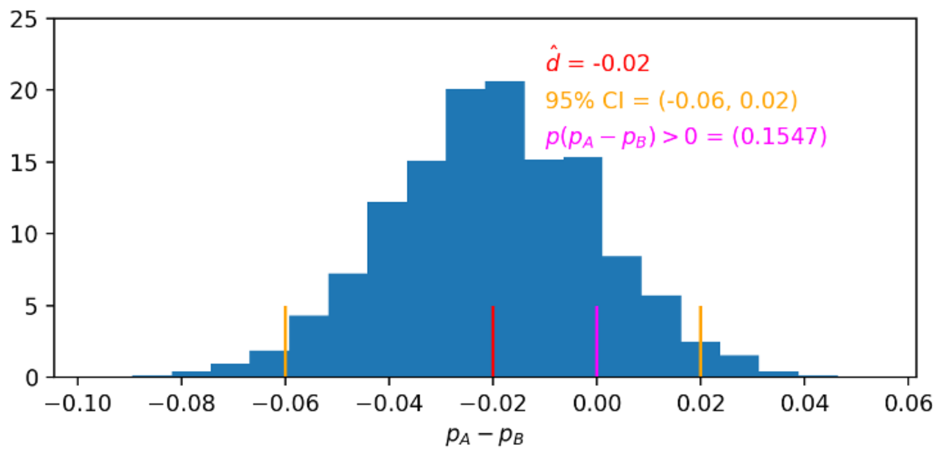
$$= \sqrt{\frac{0.70(1-0.70) + 0.72(1-0.72)}{1000}} = 0.020$$

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

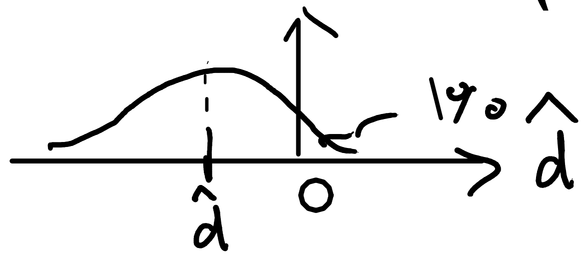
$$\Rightarrow \text{CI}: \left(\hat{\delta} - z_{\alpha/2}\hat{\sigma}_{\hat{d}} \;,\; \hat{\delta} + z_{\alpha/2}\hat{\sigma}_{\hat{d}}\right)$$

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

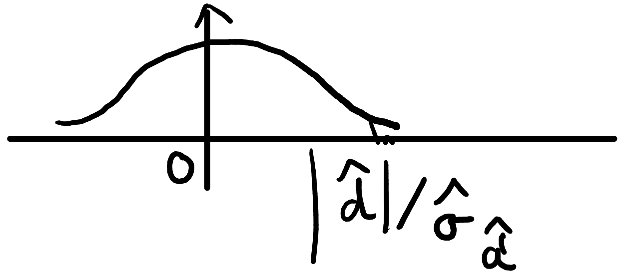
$$= (-0.06, 0.02)$$



(Sample size calculation)



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



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

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

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

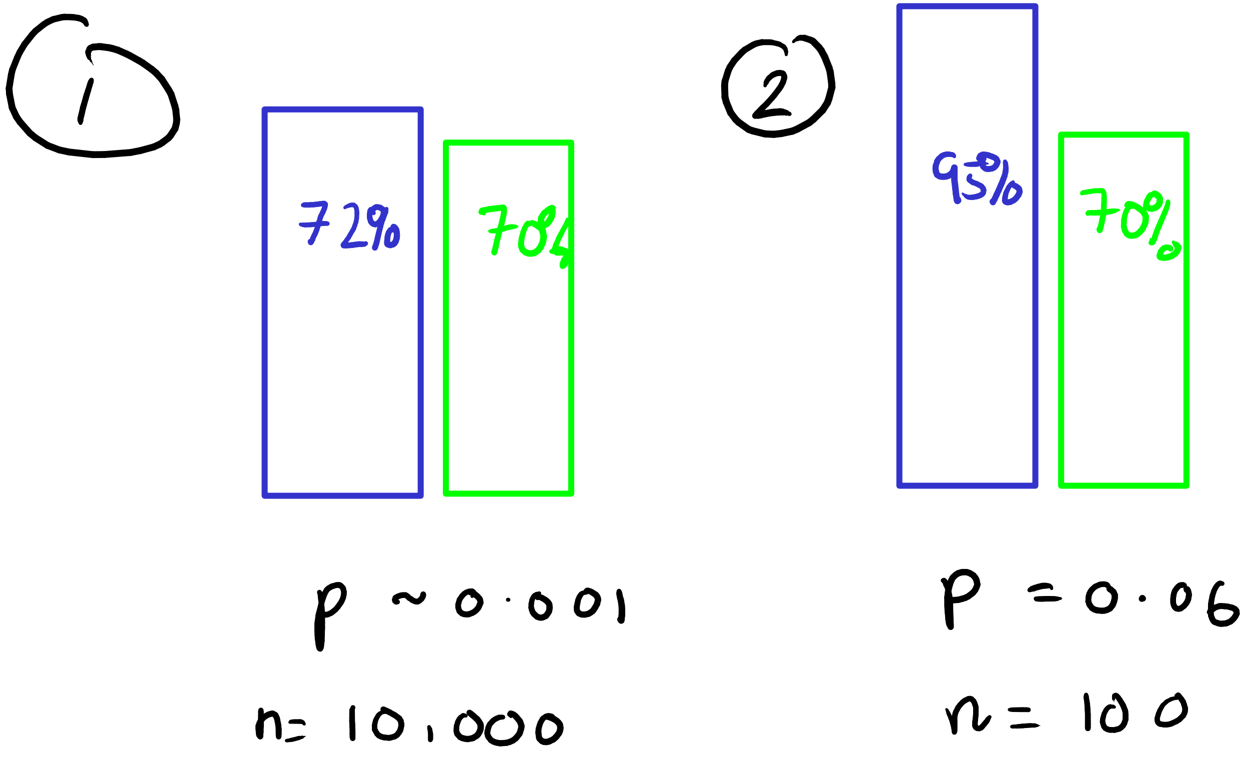
Inf2 - Foundations of Data Science:
A/B testing -
Issues in A/B testing

THE UNIVERSITY of EDINBURGH informatics

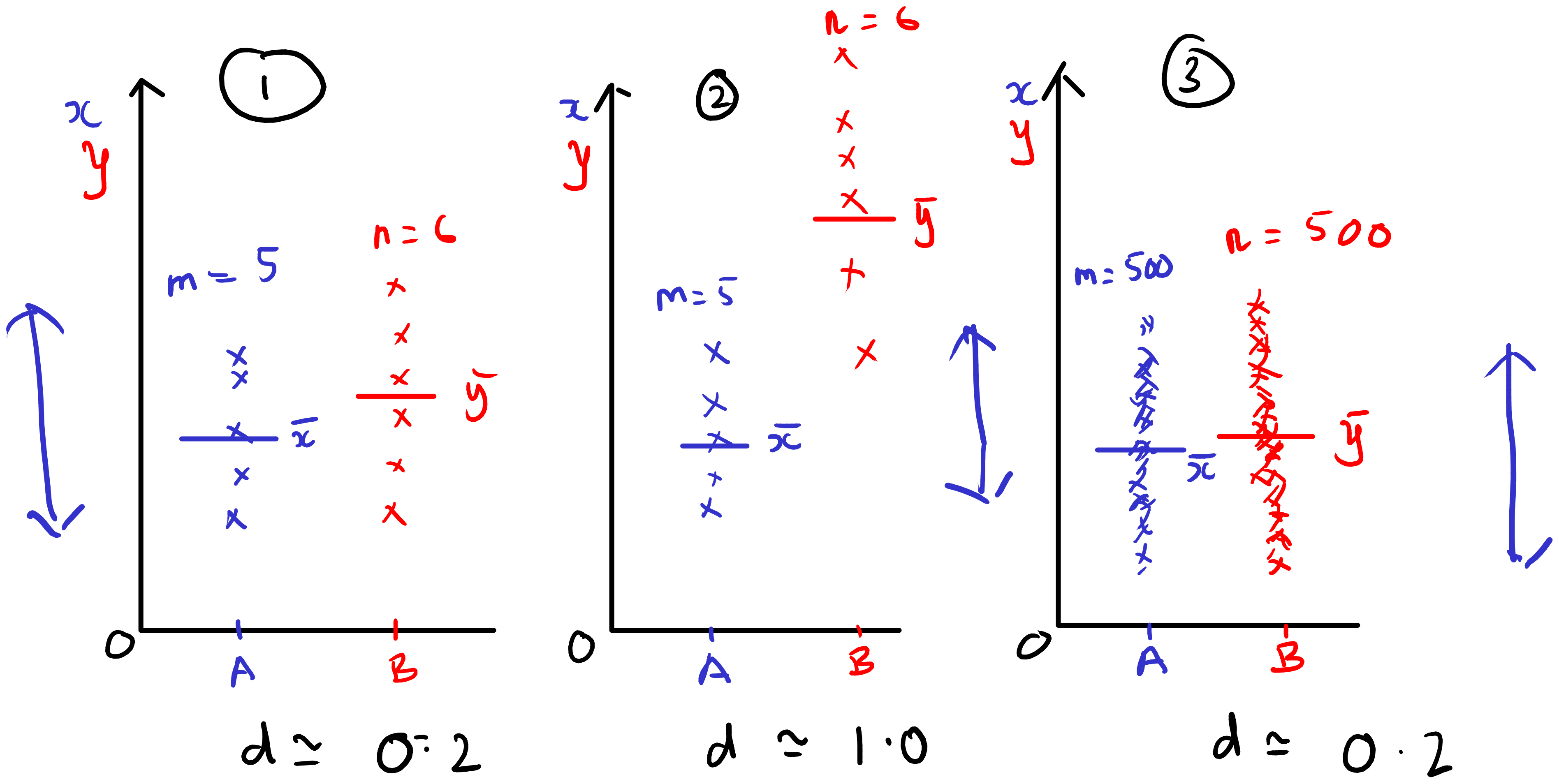
FOUNDATIONS OF DATA SCIENCE

Statistical versus practical significance

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



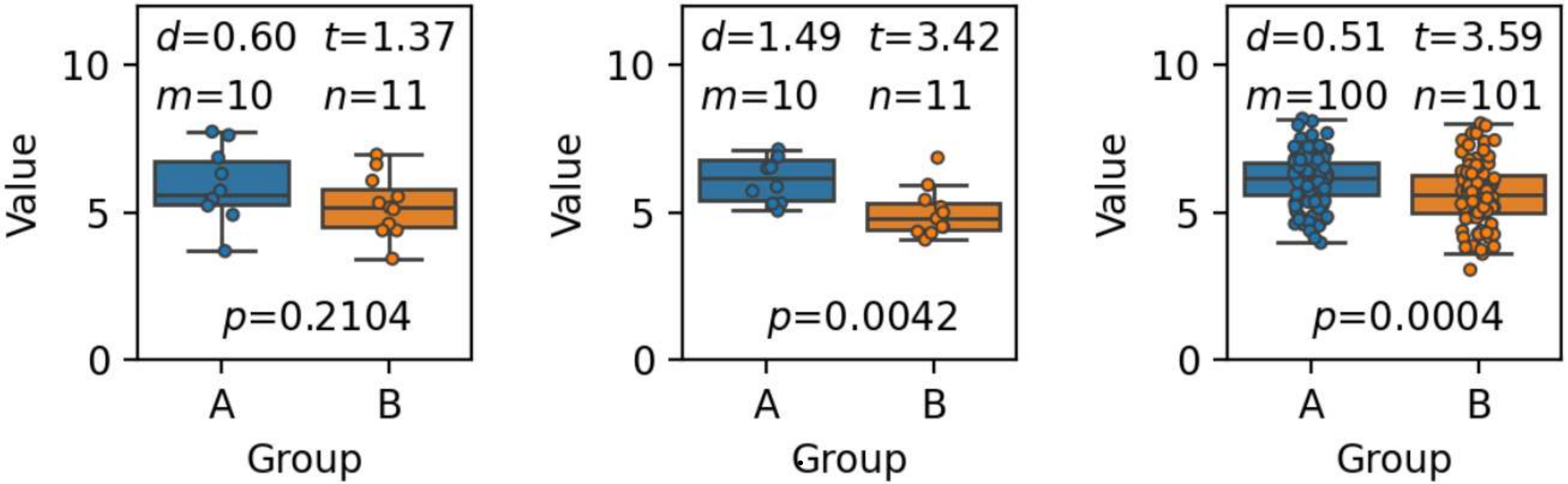
Effect size - Cohen's d



$$d = \frac{\bar{x} - \bar{y}}{s}$$

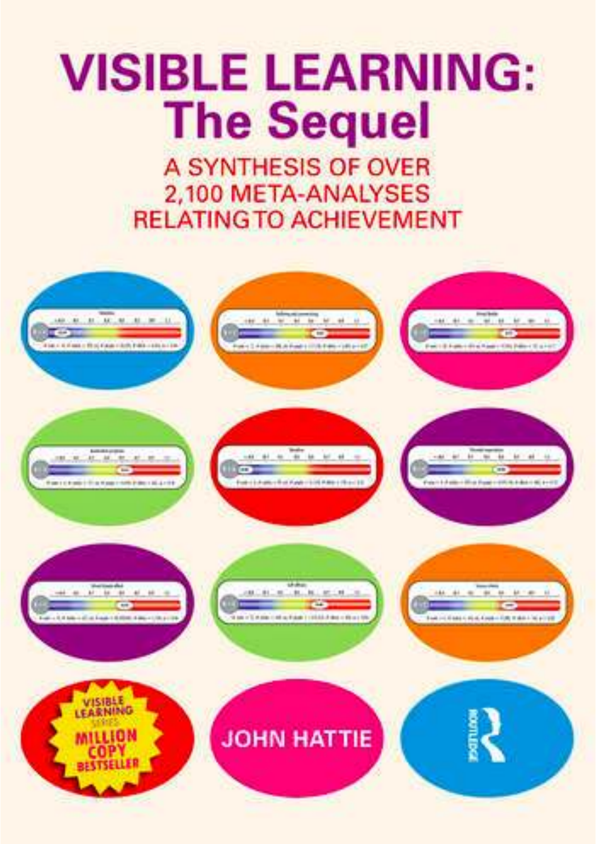
$$s = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}$$

Interpretation of Cohen's d
(Cohen (1988), Sawilowsky (2009)



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

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/

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

Inf2 - Foundations of Data Science:
A/B testing -
Comparing paired numeric samples

THE UNIVERSITY of EDINBURGH
informatics

FOUNDATIONS
OF
DATA
SCIENCE

A question and a paired experimental design

moz://a HACKS

Download Firefox

Search Mozilla Hacks

Comparing Browser Page Load Time: An Introduction to Methodology

By Dominik Strohmeier, Peter Dolanjski

Posted on November 20, 2017 in Featured Article, Firefox, Firefox Releases, and Performance

On blog.mozilla.org, we shared results of a speed comparison study to show how fast Firefox Quantum with Tracking Protection enabled is compared to other browsers. While the blog post there focuses on the results and the speed benefits that Tracking Protection can deliver to users even outside of Private Browsing, we also wanted to share some insights into the methodology behind these page load time comparison studies and benchmarks for different browsers.

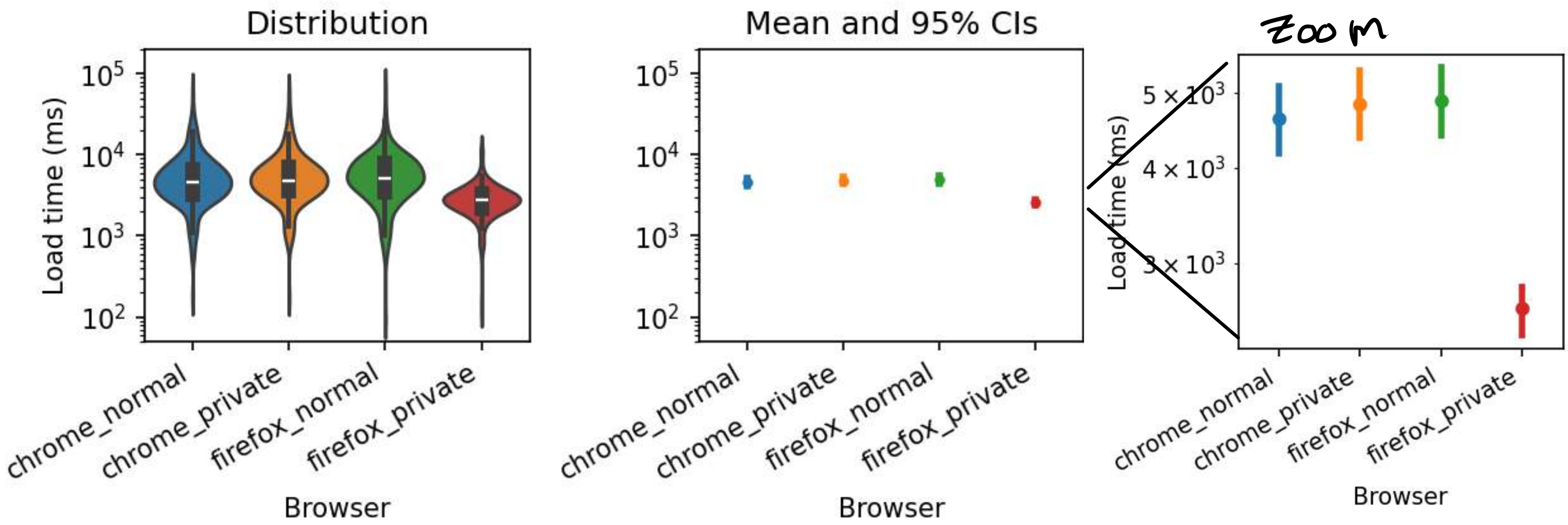
A general approach to comparing performance across browsers

https://edin.ac/3Cfl2ag

Browser	chrome_normal	chrome_private	firefox_normal	firefox_pr
Domain				
http://Abcnews.go.com	3.650247	3.618284	3.594570	3.41
http://Accuweather.com	4.381038	4.466387	4.198777	3.71
http://Adelaidenow.com.au	3.919470	3.879825	3.825883	3.58
http://Adweek.com	3.402131	3.438538	3.424099	3.26
http://Afr.com	3.646646	3.616274	3.580835	3.44

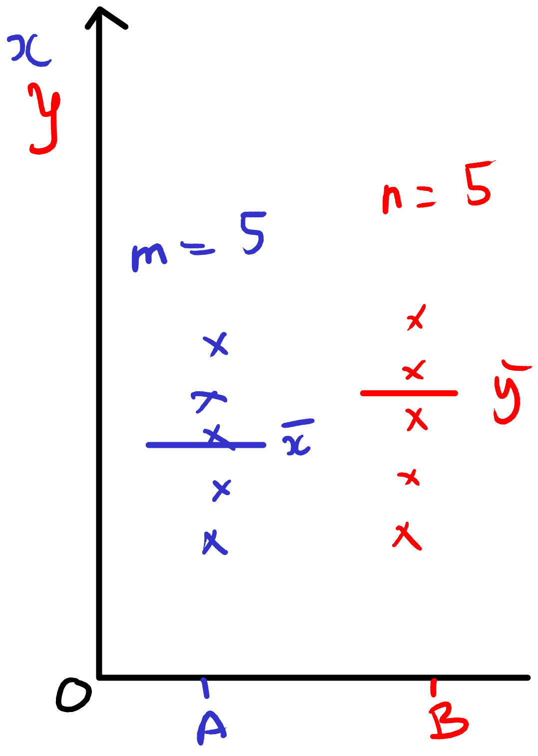
Load time of 200 popular news sites measured
10 times for each of 4 browser/configurations

Results



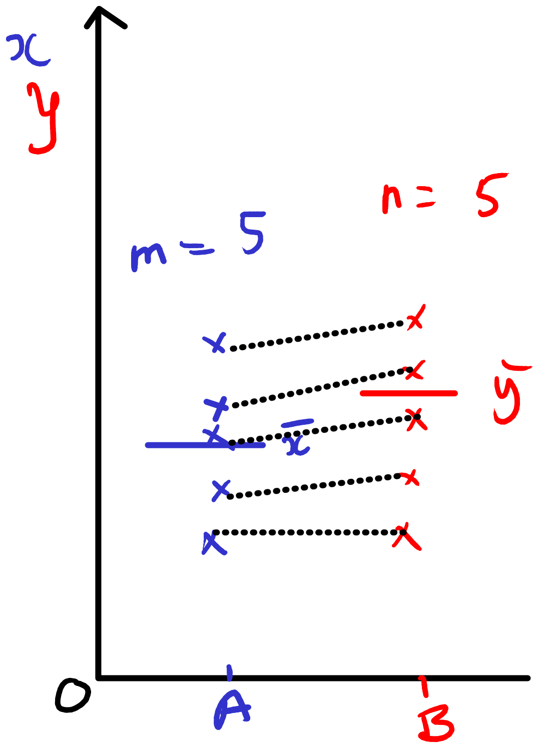
Firefox Private faster than Chrome Private?
Chrome Private slower then Chrome Normal?

Paired data



t

p - value



t

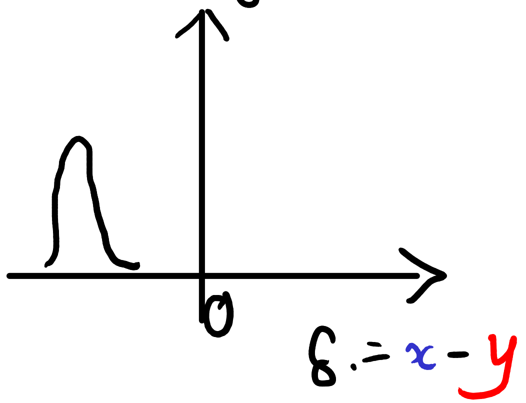
p- value

paired t-test

$$\delta_i = x_i - y_i$$

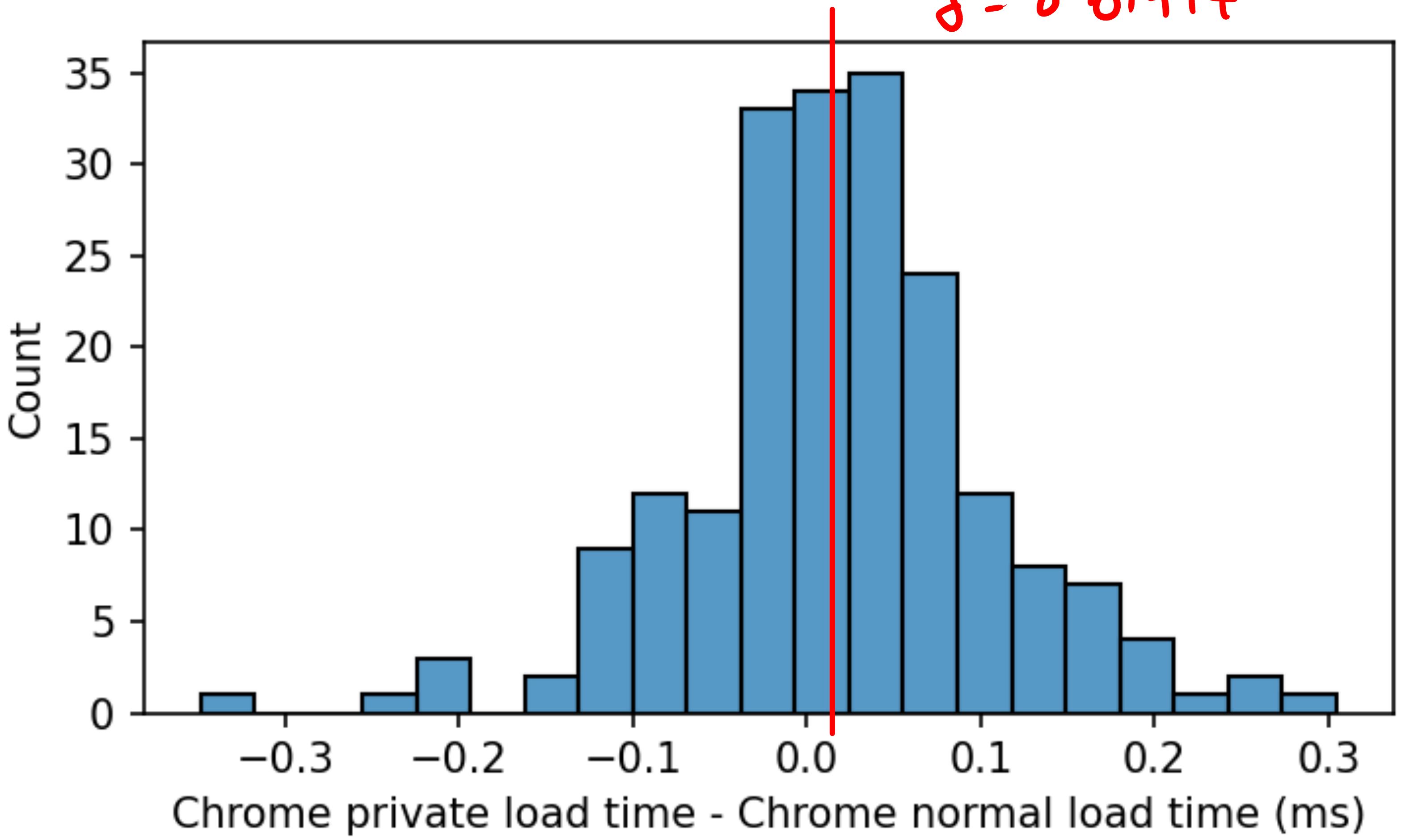
$$\hat{\sigma}_{\hat{\delta}} = \sqrt{\frac{1}{n} \sum (x_i - y_i)^2}$$

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



Paired differences

$\hat{\delta} = 0.01914$



$\hat{\sigma}_{\hat{\delta}} = 0.00628$

$\Rightarrow \quad t = \dfrac{\hat{\delta}}{\hat{\sigma}_{\hat{\delta}}} = 3.045$

p-value $= 0.0027$

Is Chrome Private slower then Chrome Normal when doing a paired test?

Summary

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

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

3. Hypothesis testing about if groups differ in means

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

4. Issues: Ethics and effect size

5. Paired design can give more statistical power

Question: Standard deviation or standard error?

What statistics should I quote to:

- A user who wants to know roughly how long they should expect to wait before reloading?

- A newspaper editor, who wants to know how long on average her journalists spent waiting for news sites to load each day (they check 100s of time a day) .