Inf2 - Foundations of Data Science: A/B testing

Announcement

Full task now available for Week 4 Workshop

Office hour starts at $10^{\underline{20}}$ today

Careers fair this afternoon!

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 numeric samples

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

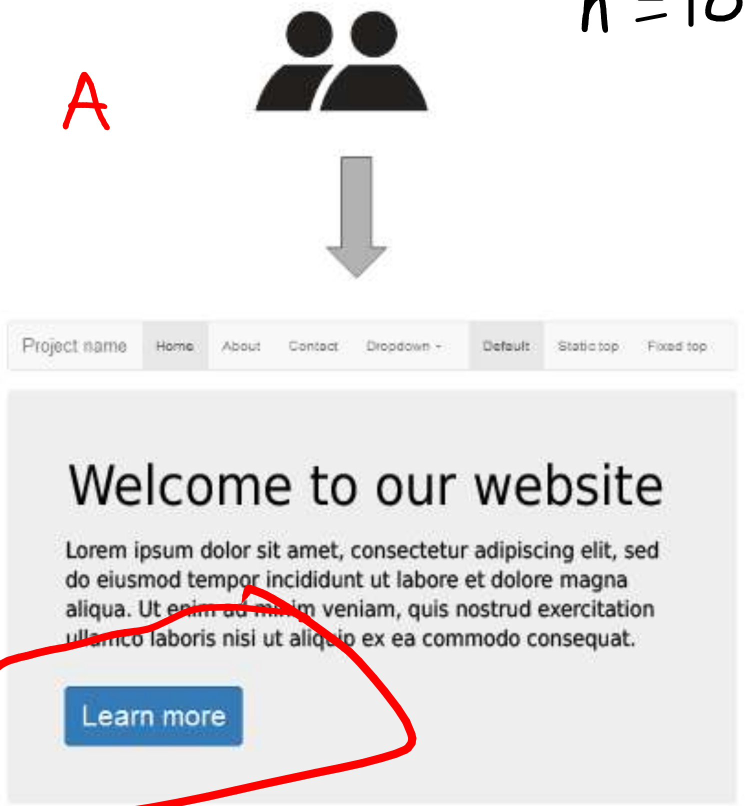
THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

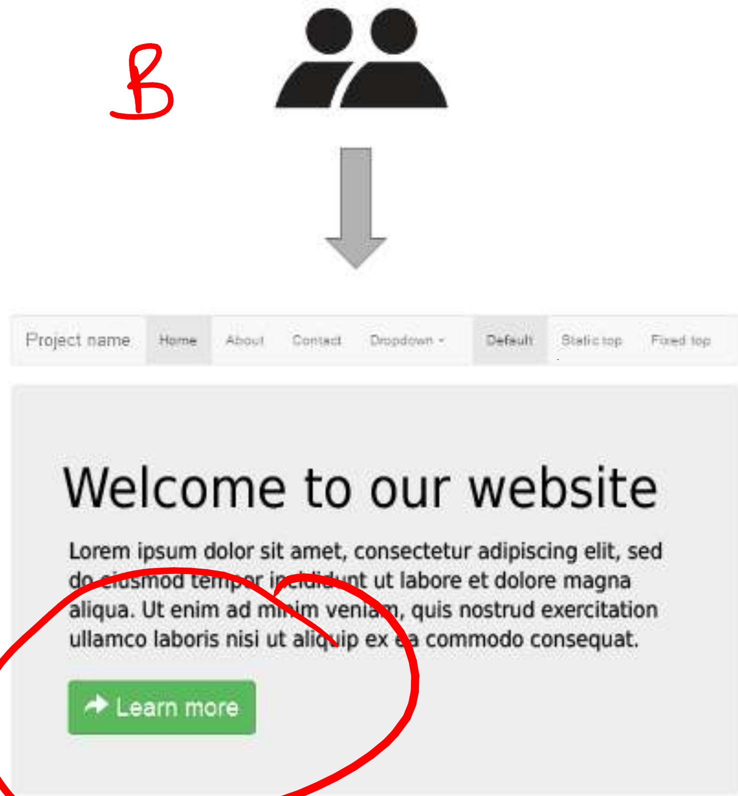
A/B Testing

$n = 1000$

A

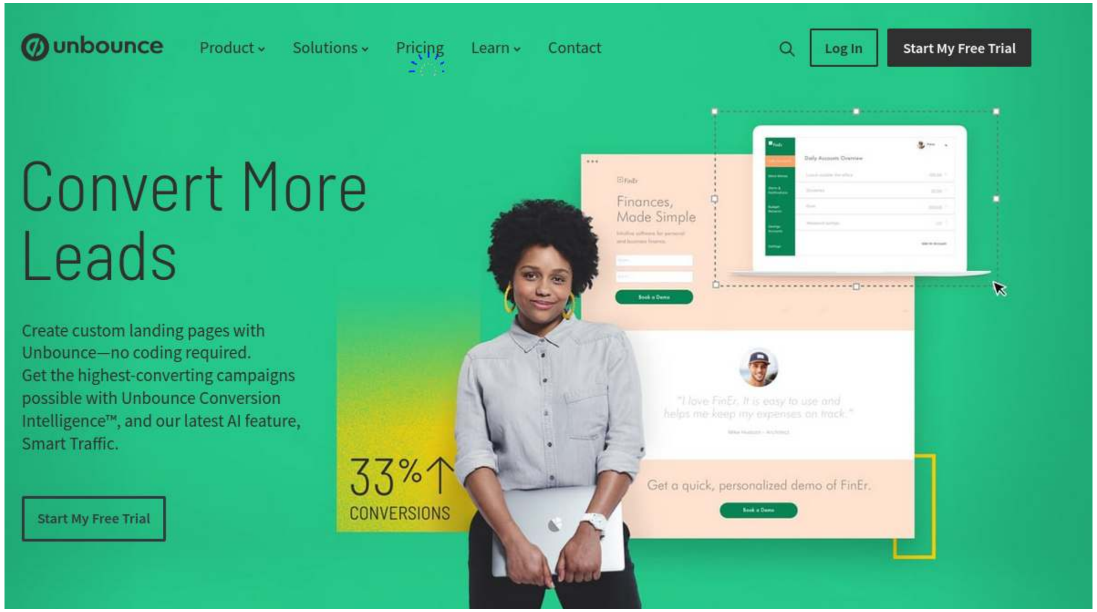


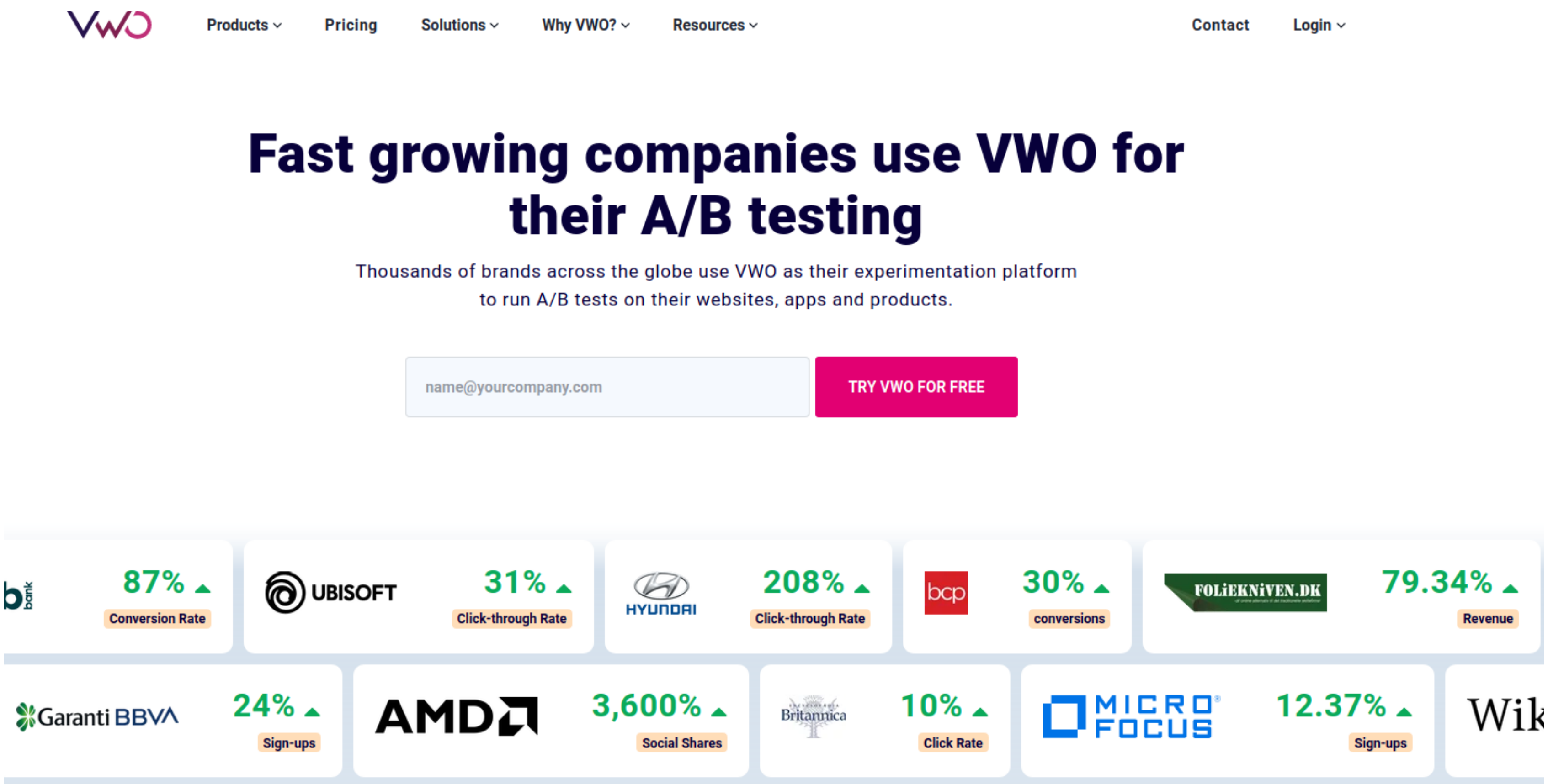
B



Click rate: ~~52 %~~ 70 % 72 %

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

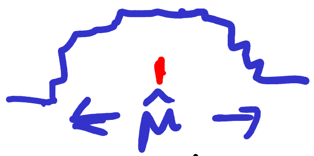

unbounce
Product
Solutions
Pricing
Learn
Contact
Log In
Start My Free Trial
Convert More Leads
Create custom landing pages with Unbounce—no coding required.
Get the highest-converting campaigns possible with Unbounce Conversion Intelligence™, and our latest AI feature, Smart Traffic.
Start My Free Trial
33%
CONVERSIONS
Finances, Made Simple
"I love FinEr. It is easy to use and helps me keep my expenses on track."
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

0. Decide underlying parameter to infer μ
1. Construct formula for estimator in terms of data $\hat{\mu} = \bar{x}$
2. Find approx. sampling distribution of estimator using bootstrap or large sample theory



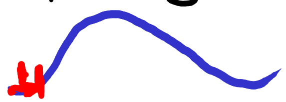
3. Return confidence interval

Hypothesis testing

0. Decide on H_0 and H_a

H_0 : Coin unbiased

1. Define test statistic in terms of data t_0
2. Find distribution of test statistic under H_0 T_0
3. Reject / not reject H_0 and find p-value



A/B testing example: Estimation approach

Parameters p_A - p_B - $\Big\}$ 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

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

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

B - # repetitions

$B = 10,000$

$$[\overbrace{1, 1, \ldots\ldots}^{700}\ \underbrace{0, \ldots, 0,}_{300}]$$

for j in 1, ..., B :

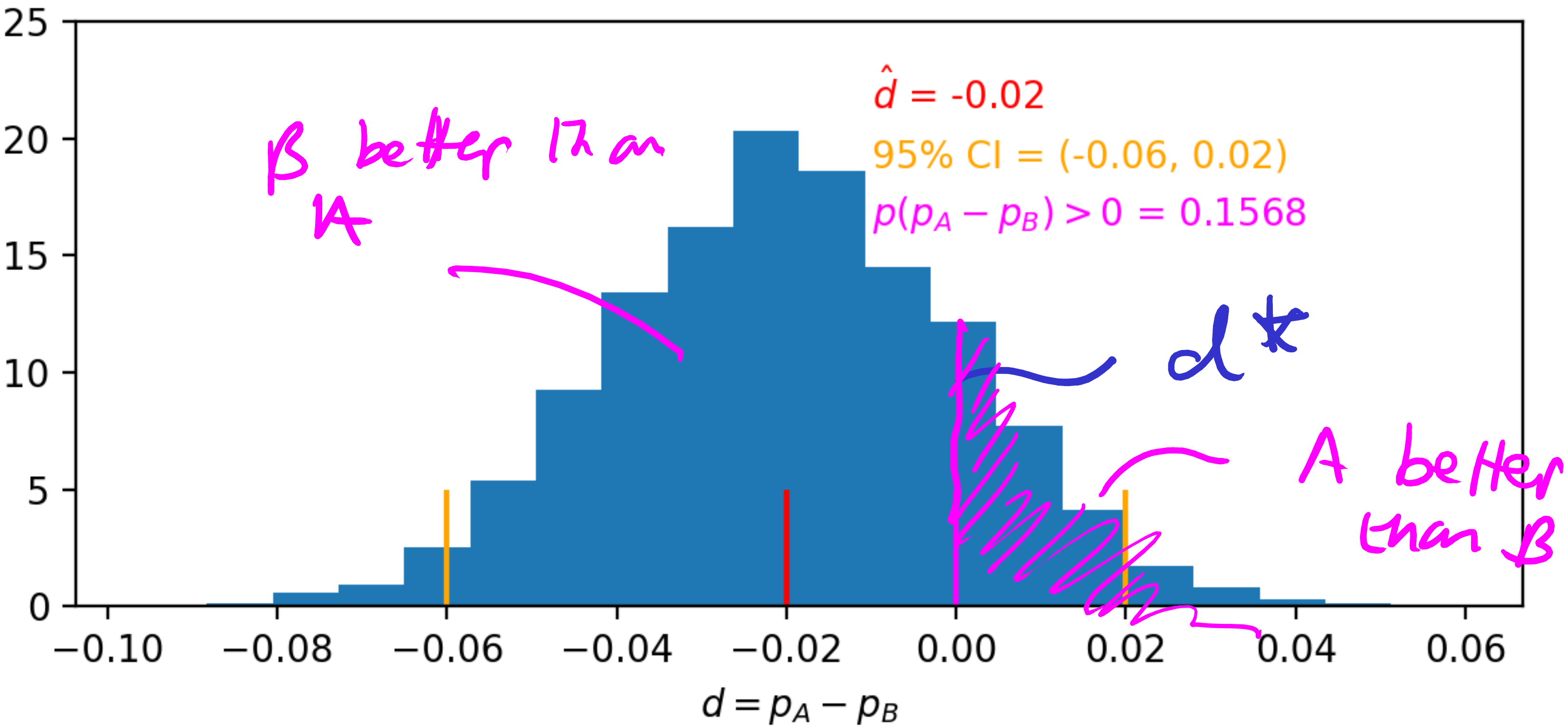
700, 695, 732

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

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

Compute quantiles, std error in estimator.

Results



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

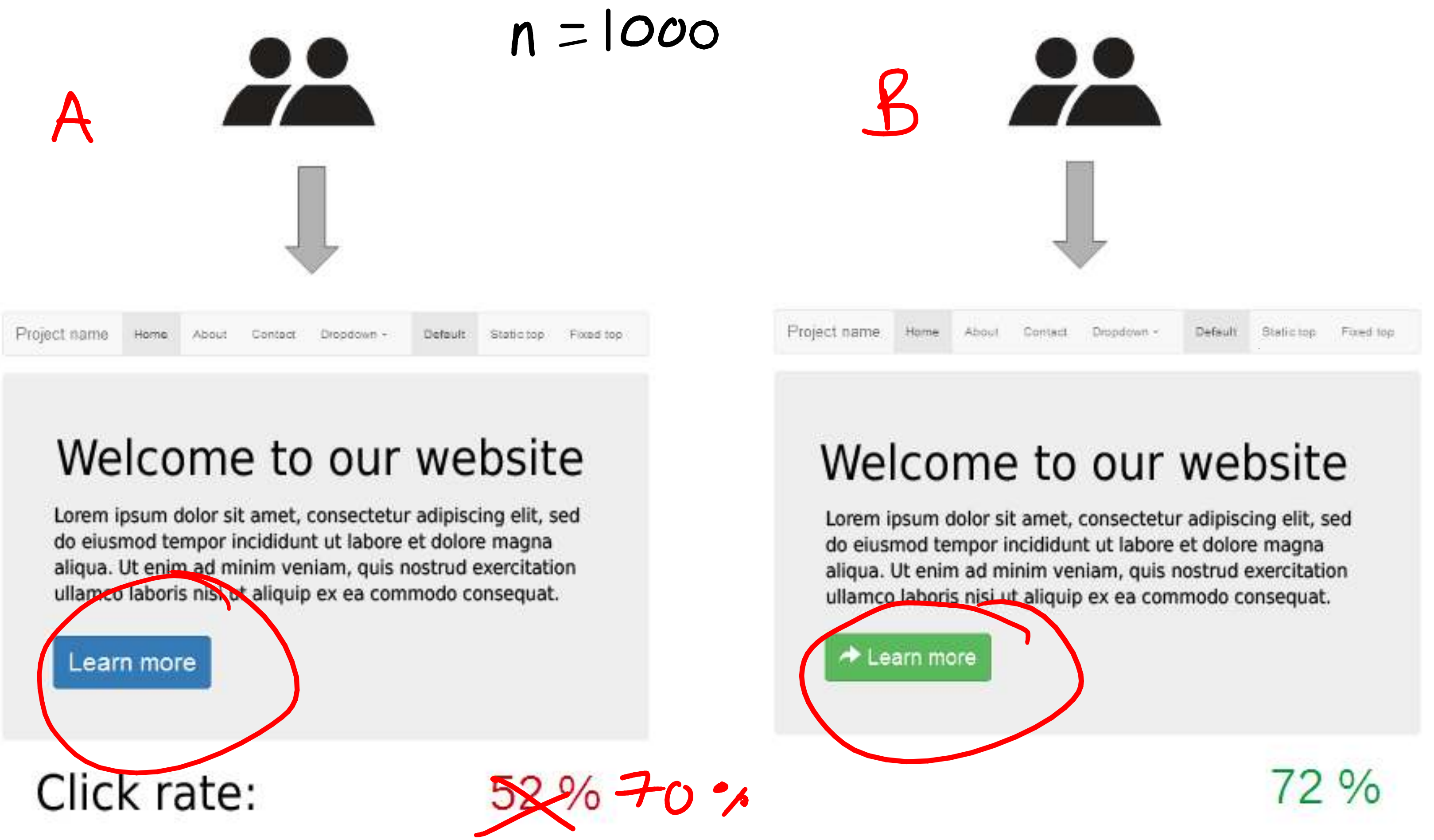
Exercise (not for the lecture)

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

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

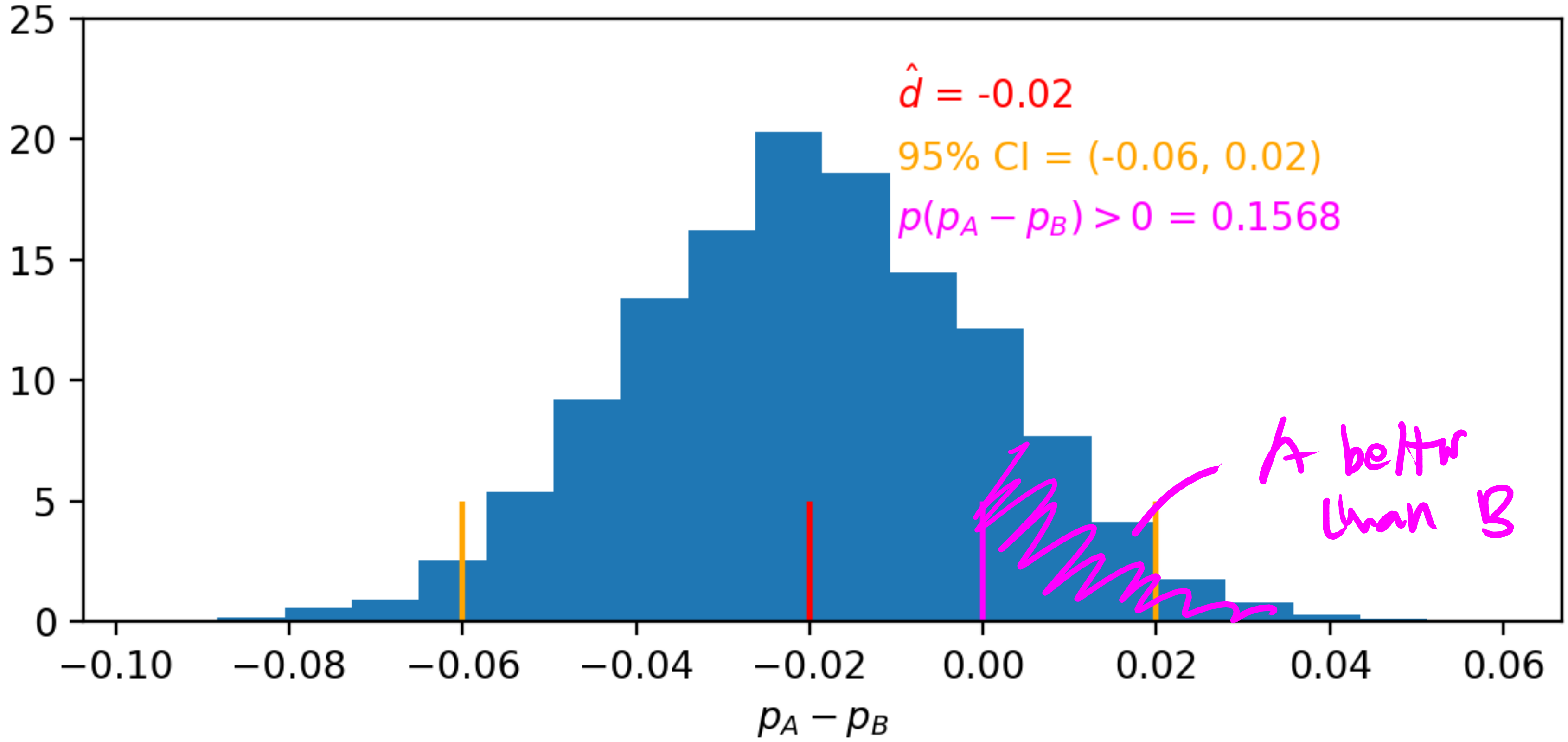
THE UNIVERSITY of EDINBURGH informatics

FOUNDATIONS OF DATA SCIENCE


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 %

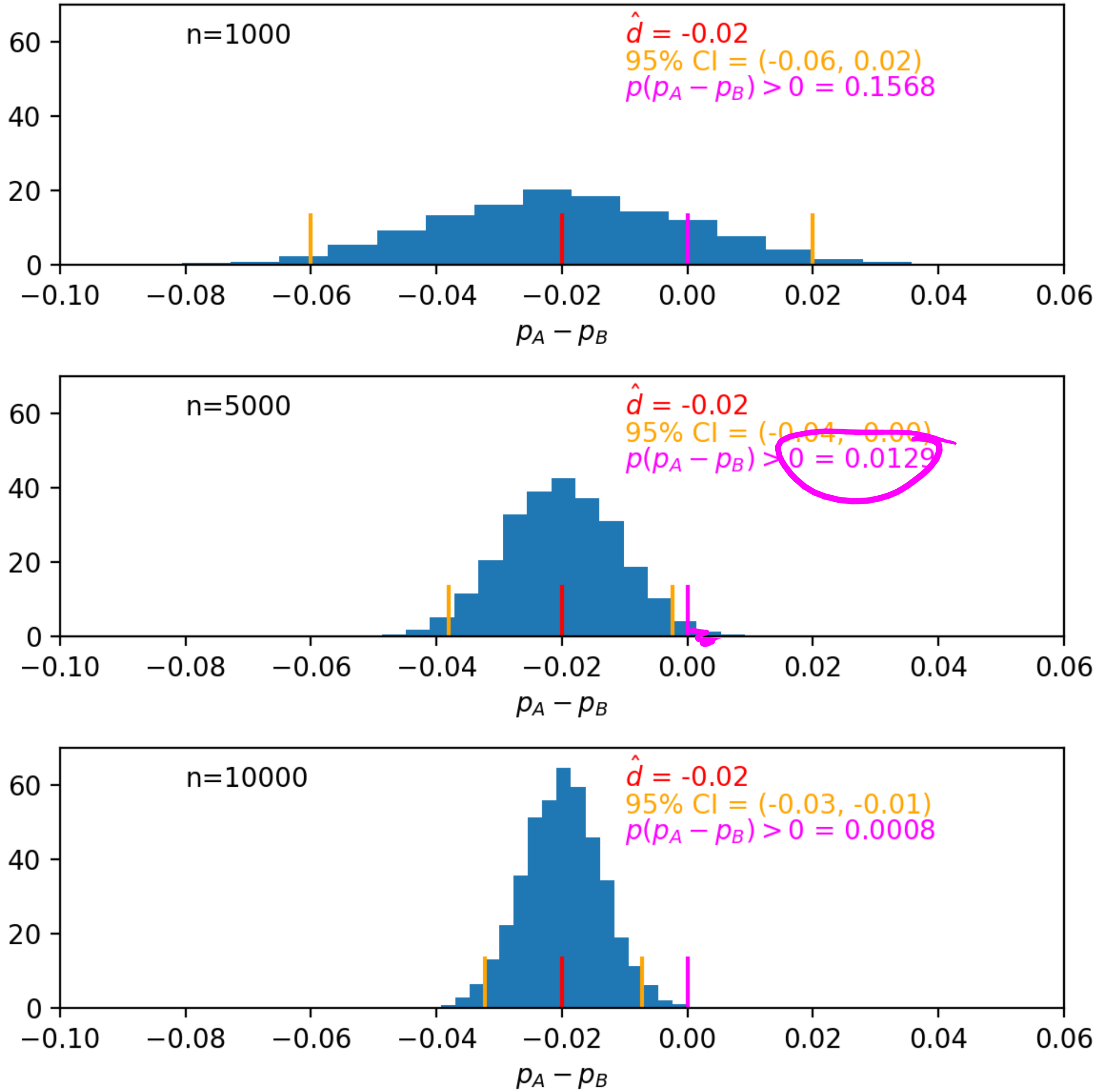
Bootstrap results

$n = 1000$



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

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 -
Large sample theory

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS
OF
DATA
SCIENCE

Standard errror of the difference

Variance of n_A $\qquad \sigma^2_{n_A} = n\hat{p}_A(1-\hat{p}_A)$

Variance of $\hat{p}_A$ $\qquad \sigma^2_{\hat{p}_A} = \dfrac{\sigma^2_{n_A}}{n^2} = \dfrac{\hat{p}_A(1-\hat{p}_A)}{n}$

" of $\hat{p}_B$ $\qquad \sigma^2_{\hat{p}_B} = \dfrac{\hat{p}_B(1-\hat{p}_B)}{n}$

Variance of $\hat{d}$ = Var(n_A) + Var(n_B)

$$\sigma^2_{\hat{d}} = \sigma^2_{\hat{p}_A} + \sigma^2_{\hat{p}_B}$$

$$\Rightarrow \sigma_{\hat{d}} = \sqrt{\sigma^2_{\hat{p}_A} + \sigma^2_{\hat{p}_B}}$$

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

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

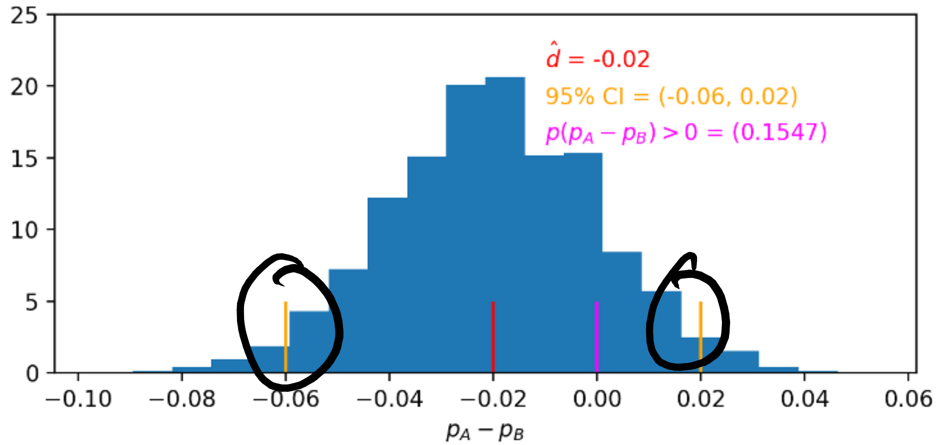
$$\hat{\sigma}_{\hat{d}} = \sqrt{\frac{0.70(1-0.70) + 0.72(1-0.72)}{1000}} = 0.020$$

$$95\% \text{ CI} \Rightarrow \alpha = 0.05 \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96$$

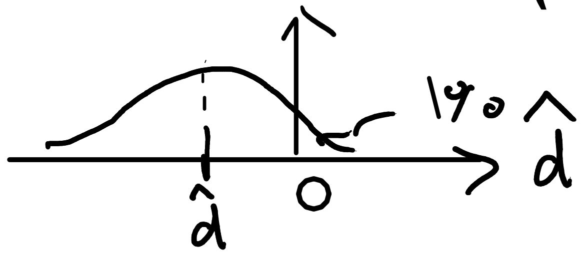
$$\Rightarrow \text{CI}: \left(\hat{d} - z_{\alpha/2}\hat{\sigma}_{\hat{d}},\ \hat{d} + z_{\alpha/2}\hat{\sigma}_{\hat{d}}\right)$$

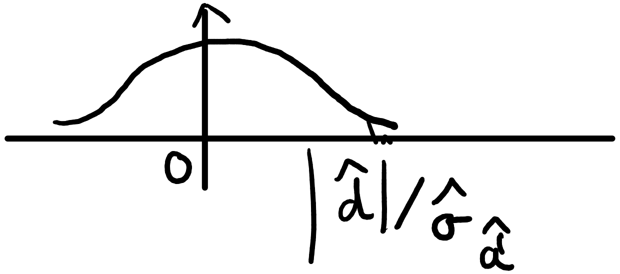
$$= -0.02 - 1.96 \times 0.020,\ 0.02 + 1.96 \times 0.02$$

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



(Sample size calculation)





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

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

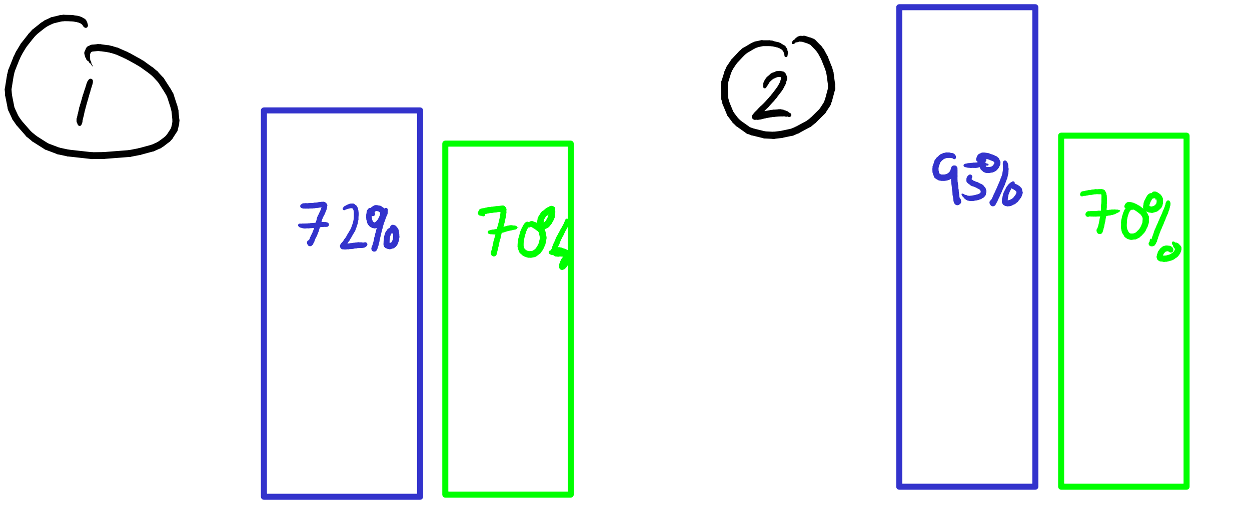
$$\frac{+\sqrt{n}\,|\hat{d}|}{\sqrt{\hat{p}_A(1-\hat{p}_A) + \hat{p}_B(1-\hat{p}_B)}} = z_{0.01}$$

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

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

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

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

THE UNIVERSITY of EDINBURGH informatics

FOUNDATIONS OF DATA SCIENCE

A question and an 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.

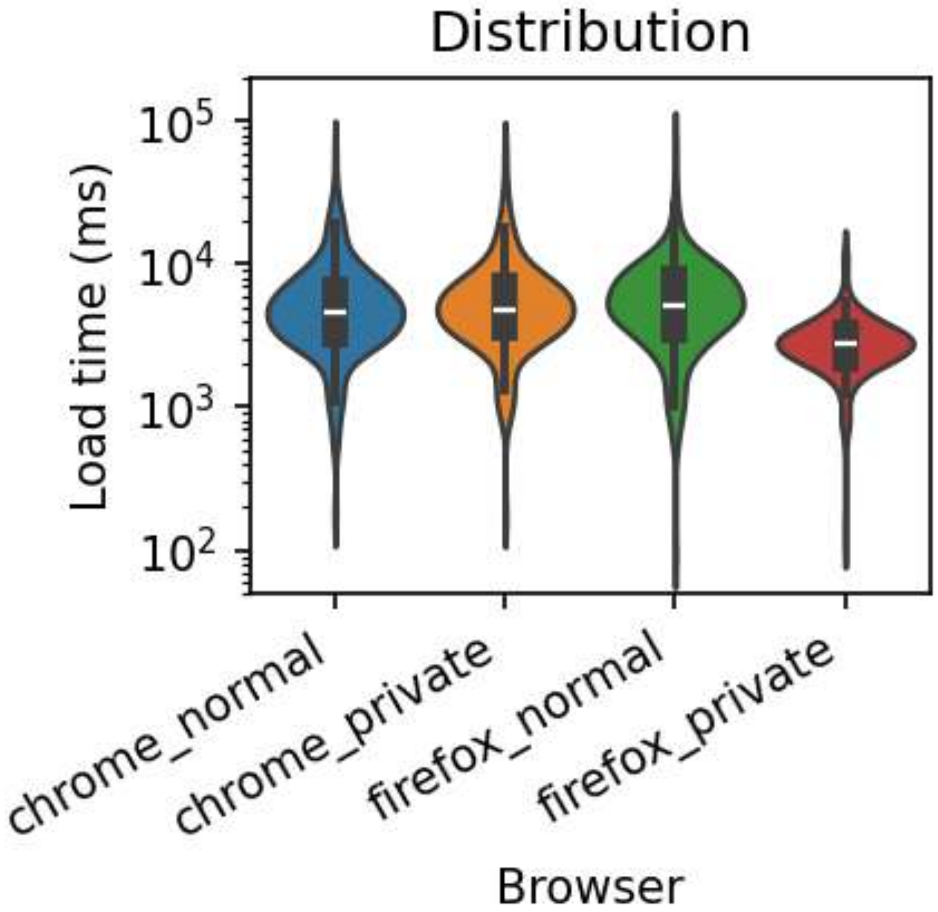
A general approach to comparing performance across browsers

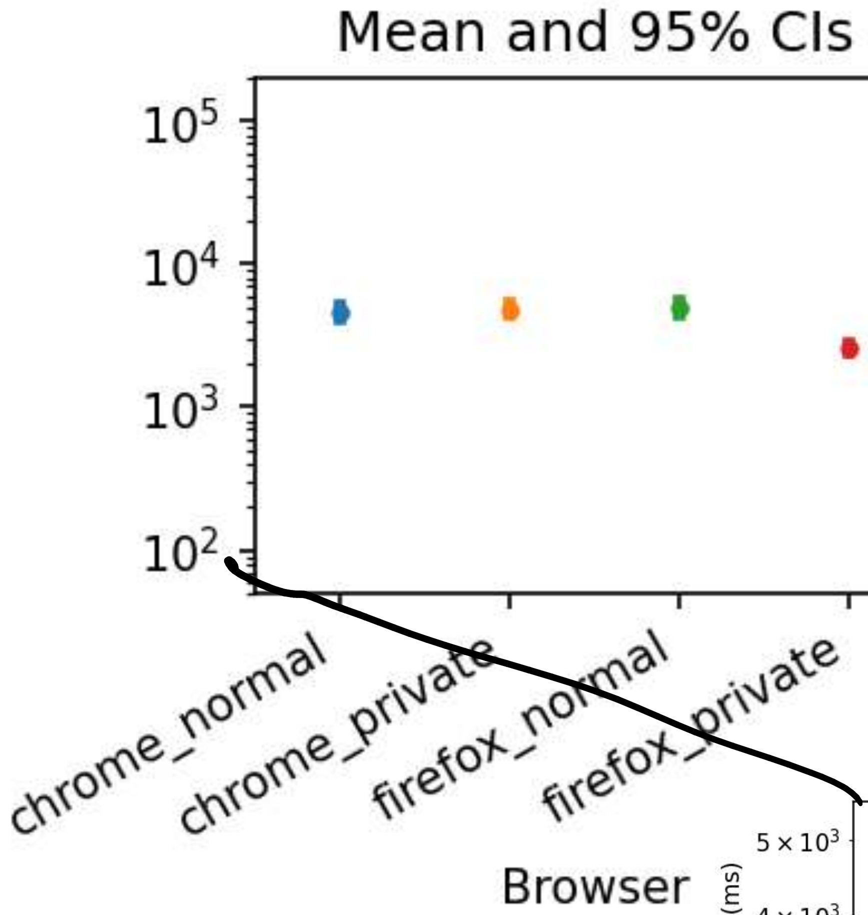
Load time of 200 popular news sites measured 10 times for each of 4 browser/configurations

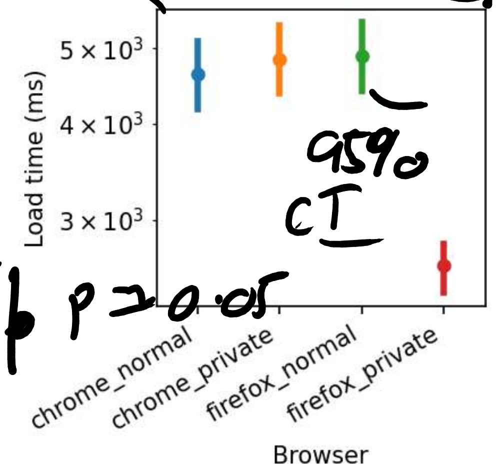
https://edin.ac/3Cfl2ag

Results

$$\frac{1}{\sqrt{200}}$$







95% CI

Firefox private faster than chome private?

Chrome private slower then chrome normal?

$p \simeq 0.001$

$p \simeq 0.05$

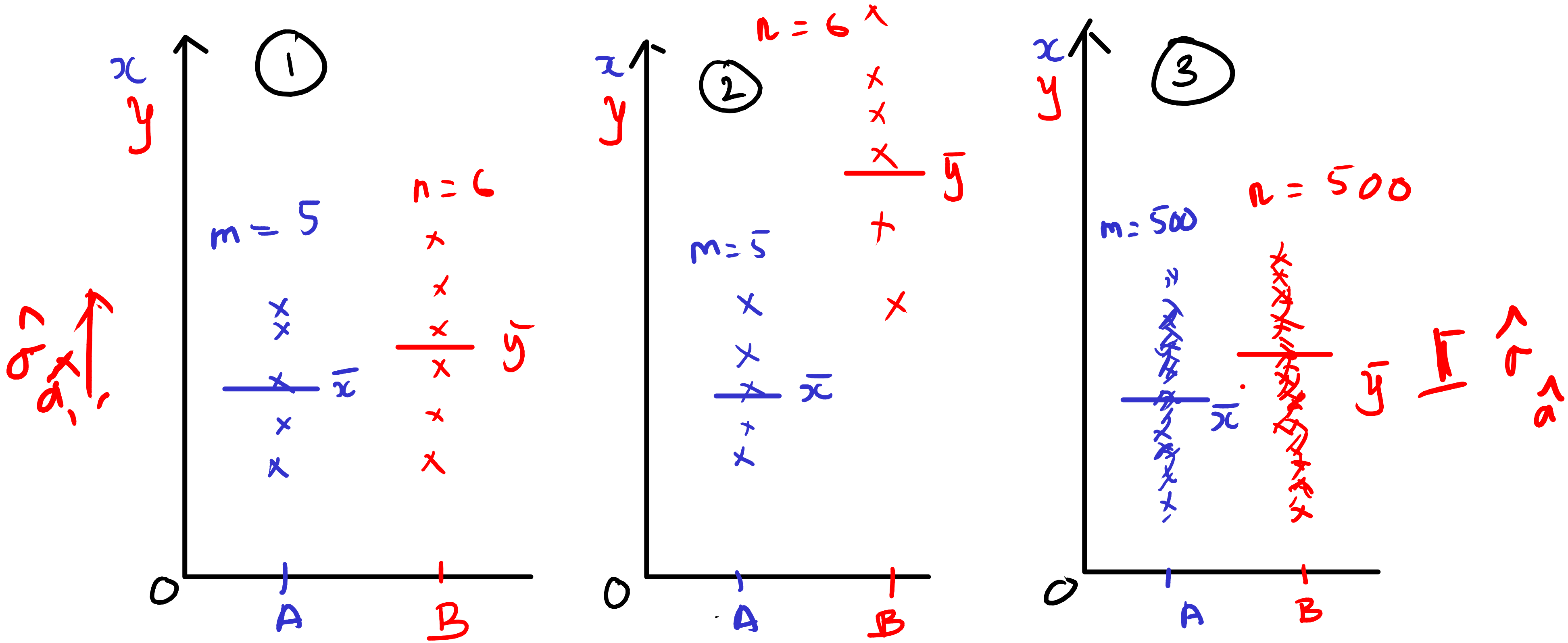
Question: Standard deviation or standard error?

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)

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



t = small   t = bigger   t = bigger

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

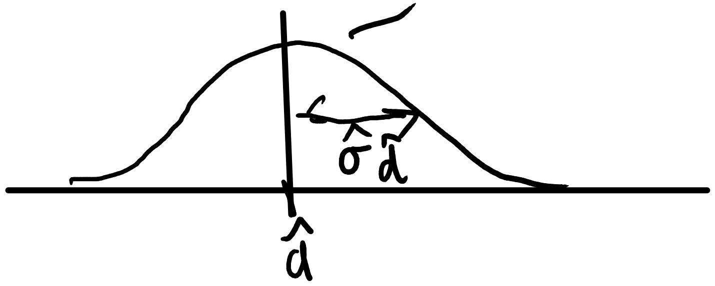
Standard error of estimator $\hat{\sigma}_{\hat{d}} = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$

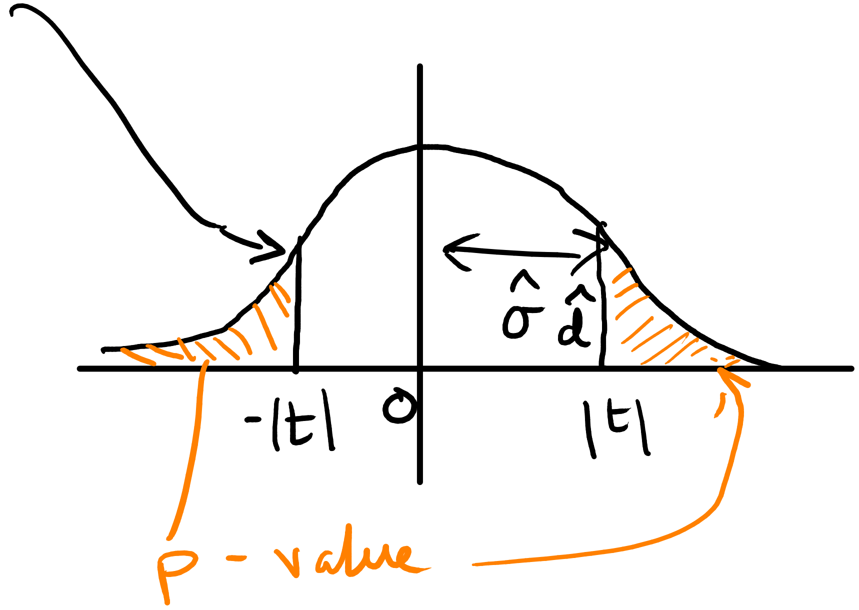
$t = \frac{\hat{d}}{\hat{\sigma}_d}$

Parameter estimation

Hypothesis test (t-test)

scaled z or t-dist

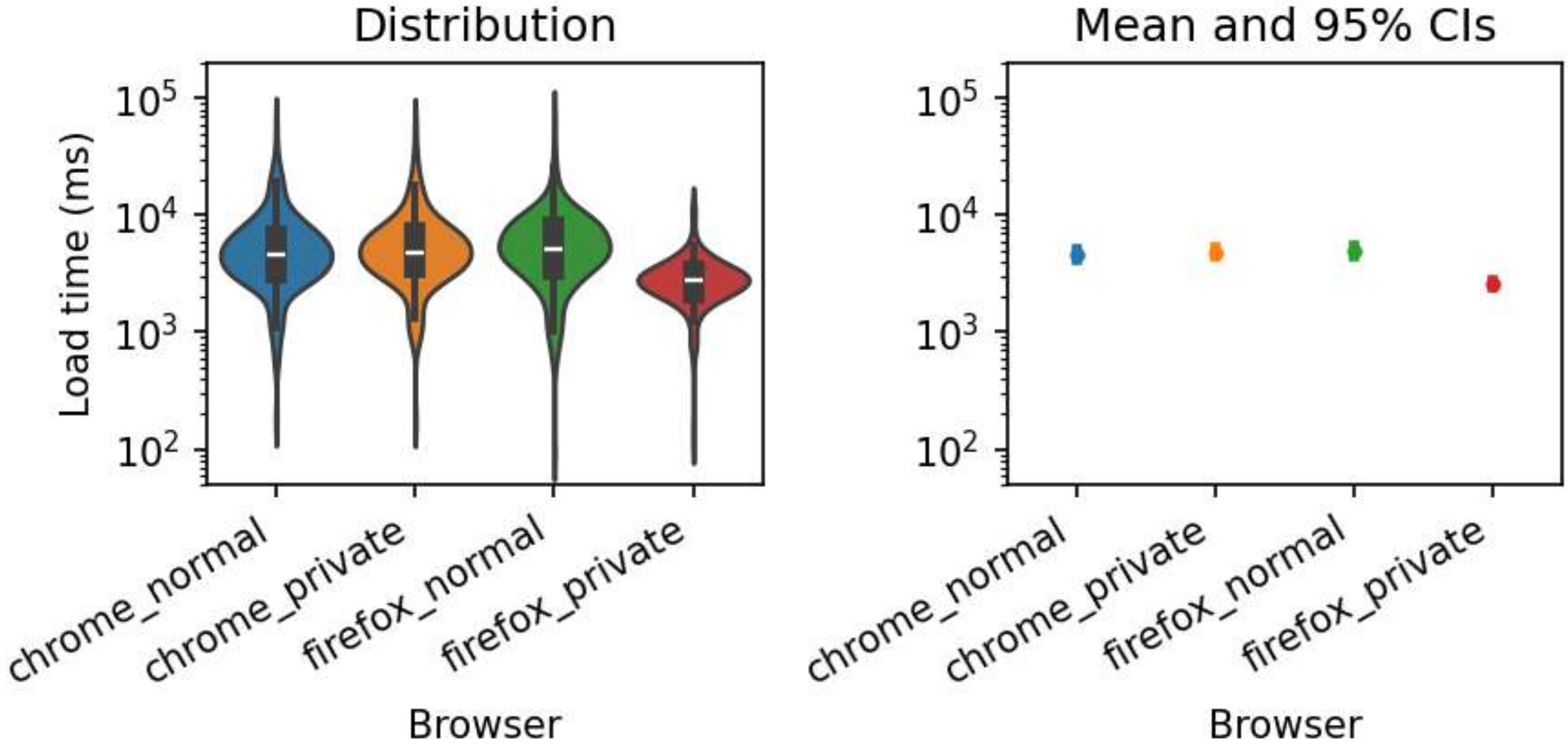




95% CI :

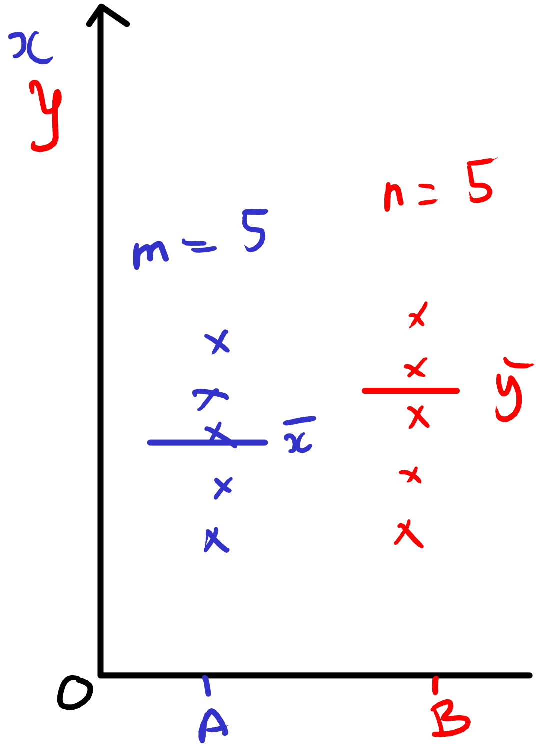
$$\left(\hat{d} - \hat{\sigma}_{\hat{d}}\, z_{0.025},\ \hat{d} + \hat{\sigma}_{\hat{d}}\, z_{0.025}\right)$$

Back to the questions

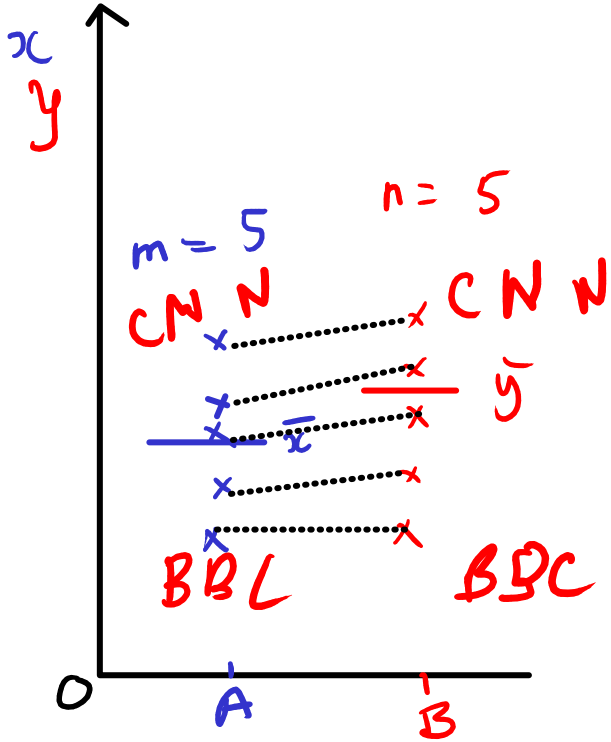


Firefox private faster than chome private?
Chrome private slower then chrome normal?

Paired data



t
p - value



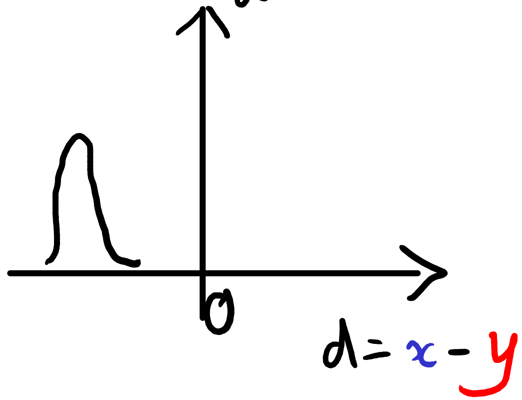
t
p- value

paired t-test

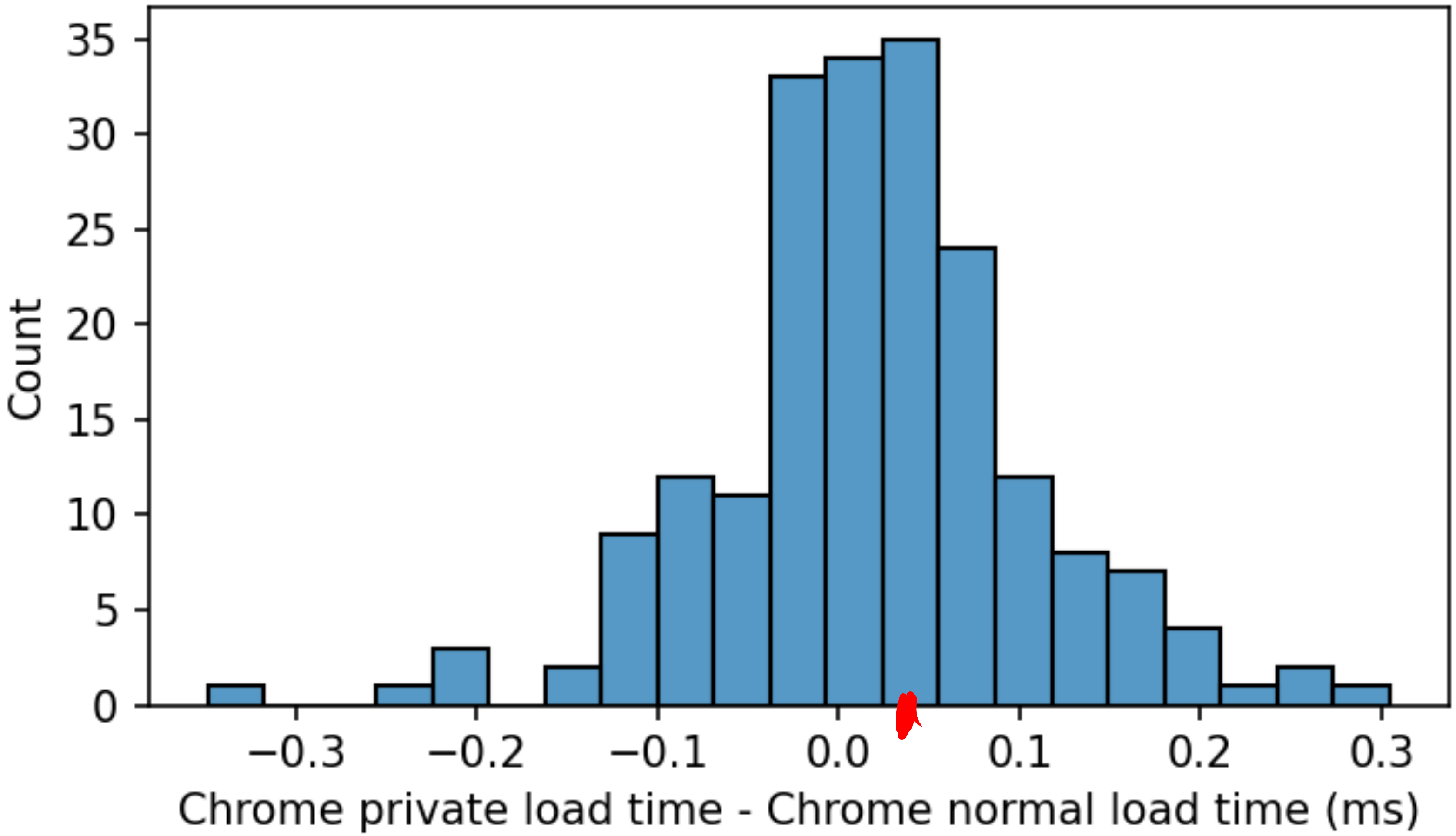
$$d_i = x_i - y_i$$

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

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

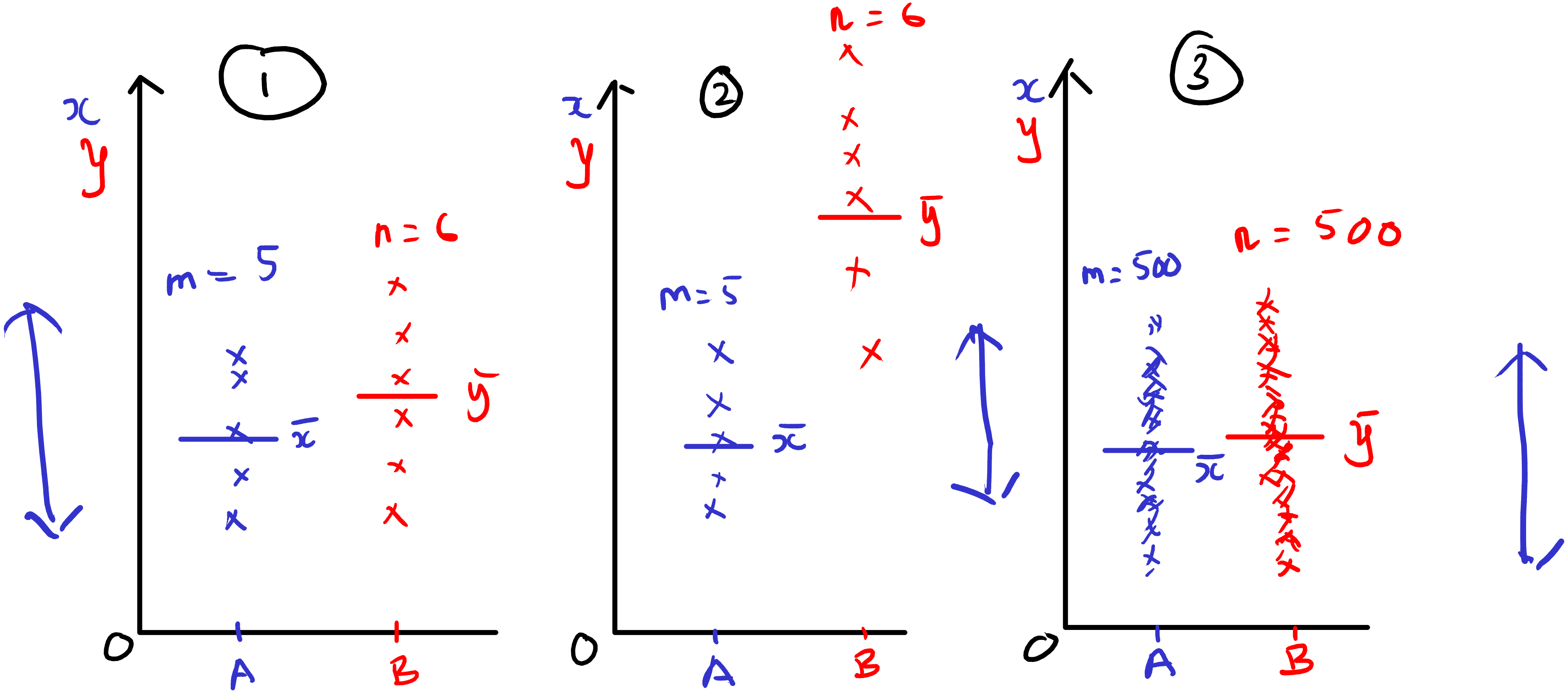


Paired differences



Could chrome private slower then chrome normal when doing a paired test?

Effect size - Cohen's d

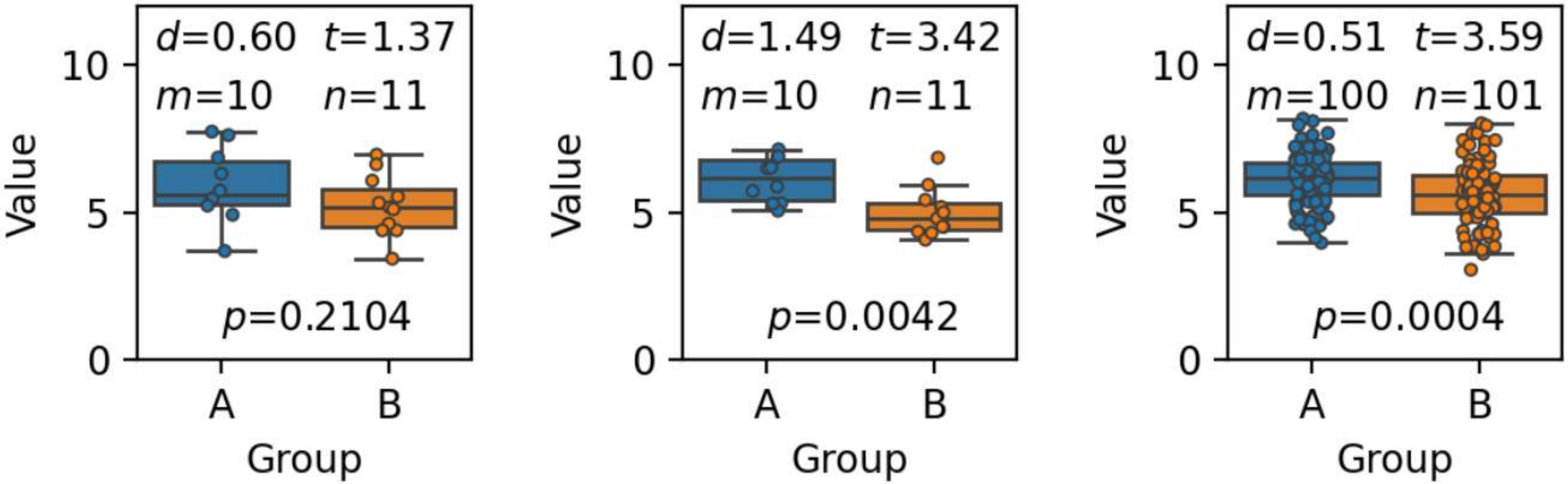


$d \simeq 0.2$ $\qquad$ $d \simeq 1.0$ $\qquad$ $d \simeq 0.2$

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

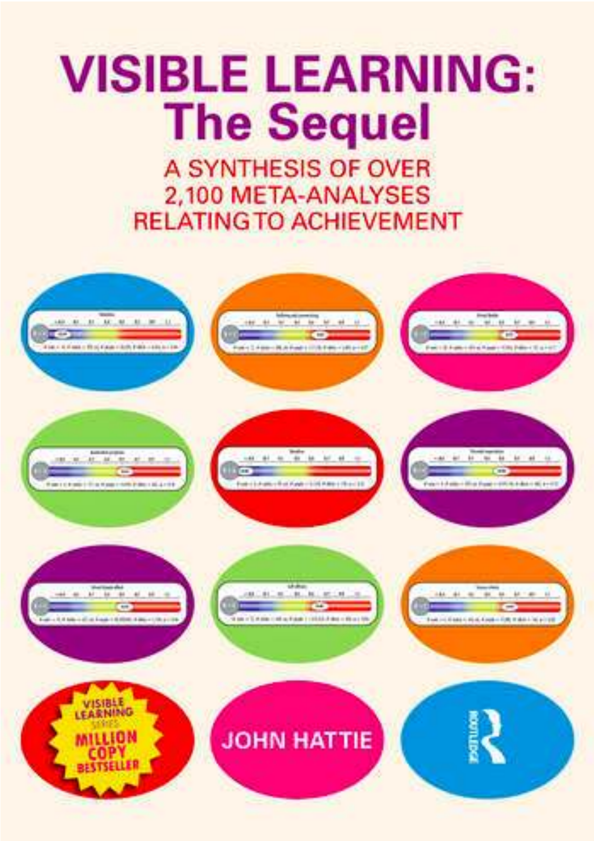
$$s = \sqrt{\frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}}$$

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/

Summary

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

2. Estimate confidence intervals between response rates in A and B, by bootstrap 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)