## **A/B Tests: The Essentials**

Lachlan Deer

Social Media and Web Analytics, Spring 2025

- 1. Explain the basic principles of an A/B Test
- 2. Analyze A/B test data to draw causal conclusions about a treatment
- 3. Determine the appropriate sample size for an experiment
- 4. Discuss challenges of shifting to an "experimentation first" company culture

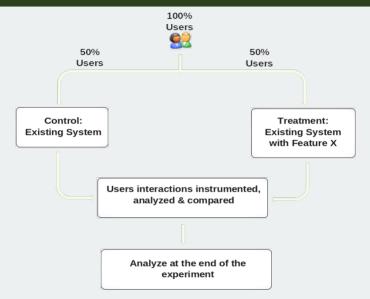
So far we've discussed:

- What makes a good research question
- The importance of **research design** and thinking through the **identification** problem to find the "right variation" to estimate casual effects
- Randomized Control Trials as a means to generate the right variation

Today: A/B tests  $\leftrightarrow$  Randomized Control Trials online!

• aka Online Controlled Experiments

### A/B Tests: The Basic Idea



## **Example: Bing Ads with Site Links**

## Shound Bing add site links to ads that allow advertisers to offer multiple destinations on an ad?

www.esurance.com/California Get Your Free Online Quote Today!	Esurance® Auto Insurance - You Could Save 28% with Esurance, Ad www.esurance.com/California Get Your Free Online Quote Today! Get a Quote - Find Discounts - An Allstate Company - Compare Rates	5
А	В	

Question: What are the pros and cons of each design?

Question: Which one created more revenue for Bing?

## **Example: Bing Search with Underlined Links**

#### Does underlining a link impact clickthrough?

bing	amazon		Q.	bing 🕨	amazon		م ا	
	Web Images Videos Maps	News More	803 🍸 Romy 🖉 🛇		Web Images Videox Maps No	eve More	803 🏋 Rarey 🔛 🗇	
	219,000,000 RESULTS Any time + Amazon com ID Official Site - Amazon		Amazon.com		210.000.000 RESULTS Any Inn - Amazon.com # Official Site - Amazo		Amazon.com	
Air www.Amazoncom. 1300 X101-16 Microson in Tellins Hugo Bulkers and Ansatury Jones, Perret Proc Cay Theory Theory and Amazon Carlos and Amazon Carlos You have reliable arcsense care to large Target Amazon Carlos The Data Market Carlos Amazon Carlos The Data Market Carlos The Data Market Carlos		Company Answars cons, loc. Is an Anner(san electronic commence consulty with headquarters in Seefle, waterington. To the largest himself based existing in the Uside Disco. Acad some interface of an order basedcore. Lot som diversitied, selfing + unvehigued.com		Ad wear-American come (2017) 2011 - Mitterers on Tenter Hage Studies and Anvance Prives, Prive Theory Come Studies) and Prives, Herrich Studies and Anvance Studies (2017) - Studies (2017) Year Inter Vehicle Areas on the studies of the Studies Neuroscience Control (2017) - American (2017) - Studies Ansatore Control (2017) - American (2017) - Studies Flue SUCT Lider Ansatore Control (2017) - Studies (2017) - Studies Flue SUCT Lider Ansatore Control (2017) - Studies (2017) - Studies Flue SUCT Lider Ansatore Control (2017) - Studies (2017) - Studies Flue SUCT Lider Ansatore Control (2017) - Studies (2017) - Studies Flue SUCT Lider Ansatore Control (2017) - Studies (2017) - Studies Flue SUCT Lider Ansatore Control (2017) - Studies (2017) - Studies Flue SUCT Lider Ansatore Control (2017) - Studies (2017) - Studies Flue SUCT Lider Ansatore Control (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) - Studies (2017) - Studies (2017) - Studies Flue SUCT (2017) -		Company Amazon com, true is an American electronic Company Amazon company with bandparters in Sauer Washington. It is the Waget Interest has de ontainer in the Usated Datas. Amazon com sauer a sau entre Eostabler, but son diversified selling +		
	Amazon com - Official Sta averatures of this data Comme while the 2003 State part activities of books magazines mark. (MOs, One stepping there for early stigger activities of books magazines, mark, (MOs, video, activities, response stituters, append a accessing, sheep, junctly		Addevac: 440 Terry Are N, Soatha, WA 90109 Founded: Ad 91, 1994 : Balleraw, WA Castomer service: +1.080-200-3321 CEO: WH Depos		Amazon.com + Official Site www.amaza.com + official site Contrast works Bill-Distribution Site Site Site Site Site Site Online response from for earth's Higger statedies of backs, magazines, most, DNDs, robert, electronics, terrorises, storens, agazine & accessions, threek greenty		Addenas: 449 Tany Ann N, Snatler, VA 99319 Ferendet: Ar 96, 1994 - Delavor, VA Customer nervice + 1880-280-3321 CCD: AP Bases	
Books Coline shopping for Books from a great coline shopping for Books from a great selection of Books Featured great selection of Uberker &		Founder: Joff Bezos Roadquarters: Souths, WA		Online sheeping for English from a great Or	indile eBooks sine shopping for Kindle Store tram a rat selection of Literature &	Founder: 2011 Discos Headquarters: Seattle, WA		
	Shop All Departments Anazos shop by department. Bite directory. This page contains links	Prime Instant Video Find, drop for and bay Prime Instant Video at Amazon.com	Stock price Bee more about AMUN • 427.63 401 4975		Amazon shop by department. She Fir develory. This page contains links Via	rime Instant Video nd, shep for and buy Prime Instant dec al Amazen.com	Stock price See more about AW2N     427.63 400	
	Models & IV         Amazon Local           DVD OPT instruct reveals 1910         Amazon Local           Mark Status 1000         Amazon Local           Decorroriss         Status 100 - 000           Other shopping from a great selection of there instructs         Shopping the tag		After Hours 427.25 - 0.30 ( 0.09%) May 20, 414 PM EDT Market Closed		Montes & IV         Amazon Local           DND 07 customs mixing 1910.99         Amazon Local           Stat 28 (c) is the runs Max         Enternantic stat is did y data website           Stat 28 (c) is the runs Max         Enternantic stat is did y data website           Electronics         Spopping           Odes stopping fore a paral selection at Decident States.         A decidy loss the max		After Hours 427.20 0.31 (0.9%) by 22, 41 Hit 121 Marker Cleane Recent post What other Is in your gloss for National Whee by Day Statement Whee Day's May Cleans In Whee Day Marker Statement S	
			Recent post span -8 the cluster of the Color is in your glass for National Wine Day? http://www.sci.us/WineDay					
	Full Store Directory Amazon shop by department, Shu directory. This page contains Teles	Outlet Outer. Online shopping for outlet deals- deal low prices on	U hars age		Amazon shop by department. Site On	utlet diel. Online shopping for outlet deals- nd low prices on	Ste more St have ago 🛐	
		Search	People also search for dee at (10+)			Search	People also search for See at (10+)	
	News about Amazon Neg continues Amazon and Perguin Rando Industry sources as Pengarita because of a contact digute	ago men black salas of its books on Amazon	ebuy eley Fightst Goods Orgh Inc. Ababa Good		News about Amazon big continue Amazon and Pengain Renders H thogarda.com - 3 minute ago biduty sources by Pengain me bocase of a contact dispute, bid	which sales of its basis on Amazon	ebary attey Pigkat Seter	

Question: Which one created more revenue for Bing?

# **1/** Working Example: Email Marketing

## An Email A/B Test

The email A/B test we will analyze was conducted by an online wine store.

	\$15 OFF EVERY	\$100 ON WINERY DIRECT WINES	Shop Winery Direct	Details	
Total Wine	Q What can we h	nelp you find today?		Search O West	tion Orange, NJ 🗸 😽
Wine $\checkmark$ Spirits $\checkmark$ Be	eer V Accessories &	More 🗸 🛛 Deals 🛛 Gift Guide			Sign In   Create Account
Home	/ Wine / Red Wine	/ Syrah/Shiraz			
Syral	h/Shiraz	Available to you (16	9) Sort by	All stores (618)	
	West Orange, NJ (0.0 miles)	Pick up in store     Mis Ship to NJ	Pickup or ship Most Po		=
	Union, NJ (7.6 miles)			•	
	River Edge, NJ (13.9 miles)	L L	Å_`	••	
	Norwalk, CT (49.7 miles)		1		
	Milford, CT (69.9 miles)		11	Jan	
View	more   Select all	_		-	
Dese	dect all	Molly Dooker Shiraz The Boxer	Yellow Tail Shiraz	Jam Jar Sweet Shiraz	
Cate	agory 🗸	750mi	1.54	750mi	
Wine	e Varietal & Type 🛛 🛩	\$ 26.97	\$ 11.47	\$ 7.46	IIII
Price	e Range 🖌 🖌 🗸			\$ 8.29 per bottle	
Bran	nd 🛩	ADD TO CART	ADD TO CART	ADD TO CART	
Cour	ntry/State 👻	SAVE TO LIST	SAVE TO LIST	SAVE TO LIST	

Source: Total Wine & More

Test setting: email to retailer email list

Unit: email address

Treatments: email version A, email version B, holdout

Reponse: open, click on link and 1-month purchase (\$)

Selection: all active customers

Assignment: randomly assigned (1/3 each)

## Loading & Inspecting the Data

Rows: 123,988

Columns: 14

\$ user id <dbl> 1000001, 1000002, 1000003, 1000004, 1000005, 1000006, 10000~ \$ cpgn id <chr> "1901Email", "1900Email", "1900Email",","1900Email",","1900Email",","1900Email",",","1900Email",", \$ group <chr> "ctrl", "email B", "email A", "email A", "email A", "email ~ \$ email <lp><lgl> FALSE. TRUE. TRUE. TRUE. TRUE. TRUE. TRUE. TRUE. TRUE. TRUE. \$ open <dbl> 0. 1. 1. 1. 1. 1. 1. 0. 1. 0. 0. 0. 0. 0. 1. 0. 1. 1. 1. 0.~ \$ click \$ purch <dbl> 0.00. 0.00. 200.51. 0.00. 158.30. 0.00. 26.52. 0.00. 0.00. ~ \$ chard <dbl> 0.00. 0.00. 516.39. 0.00. 426.53. 0.00. 0.00. 0.00. 0.00. 0~ \$ sav blanc <dbl> 0.00, 0.00, 0.00, 0.00, 1222.48, 0.00, 0.00, 0.00, 0.00, 0.~ \$ syrah <dbl> 33.94, 16.23, 16.63, 0.00, 0.00, 0.00, 124.31, 32.12, 148.5~ \$ cab <dbl> 0.00, 76.31, 0.00, 41.21, 0.00, 0.00, 58.19, 62.67, 0.00, 0~ \$ past\_purch <dbl> 33.94, 92.54, 533.02, 41.21, 1649.01, 0.00, 182.50, 94.79, ~ \$ days since <dbl> 119. 60. 9. 195. 48. 149. 118. 125. 100. 50. 192. 27. 41. 4~ \$ visits <dbl> 11, 3, 9, 6, 9, 6, 8, 7, 7, 6, 0, 4, 9, 8, 6, 6, 5, 7, 7, 9~

## Variables associated with the Test

#### Treatment indicator ( $T_i$ )

• Which (randomized) treatment was received

#### Outcomes (Y<sub>i</sub>)

• Outcome(s) measured for each customer, i.e. the outcome variable

#### **Baseline variables (** $Z_i$ **)**

- Other stuff we know about customers **prior** to the randomization
- · Sometimes called "pre-randomization covariates" or "observables"

**Question:** For each variable in the dataset, which one of these categories does it fall into?

## 2/ Analysis of A/B tests

## **The First Question**

What is the first question you should ask about an A/B test?

#### What is the first question you should ask about an A/B test?

Did the treatment affect the response?

Was the randomization done correctly?

How can we check randomization with the data at hand?

**Randomization checks** confirm that the **baseline variables** are **distributed similarly** for the **treatment and control groups**.

• Also known as "Balance tests"

# A tibble:  $3 \times 8$ group days since mean visits mean past purch mean chard mean sav blanc mean <chr> <dbl> <dbl> <dbl> <dbl> <dbl> 1 ctrl 90.0 71.7 73.6 5.95 188. 73.5 2 email A 90.2 5.95 188. 72.1 3 email B 89.8 5.94 190. 74.8 71.6 # i 2 more variables: syrah mean <dbl>, cab mean <dbl>

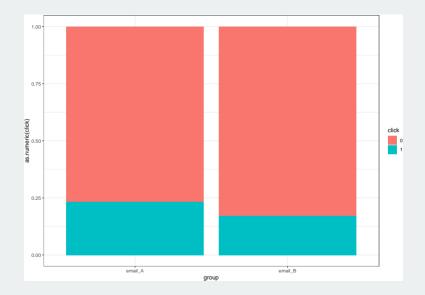
#### We can **test for balance** across treatments for each of our baseline variables:

#### Randomization seems to check out!

... onto average treatment effects

Look at the means of outcome variables between treatments:

Question: What differences do you observe?



## Does email A have higher open rate than B?

## Does email A have higher open rate than B?

## Does email A have higher click rate than B?

## Doing it all at once with regression

#	A tibble: 3 :	x 5			
	term	estimate	std.error	statistic	p.value
	<chr></chr>	<dbl></dbl>	<dbl></dbl>	<dbl></dbl>	<dbl></dbl>
1	(Intercept)	Θ	0.00127	Θ	1.00
2	groupemail_A	0.132	0.00179	73.5	Θ
3	groupemail_B	0.0934	0.00179	52.1	0

## Does email A lead to different click thru than B?

```
Linear hypothesis test
```

```
Hypothesis:
groupemail A - groupemail B = 0
Model 1: restricted model
Model 2: click ~ group
 Res.Df RSS Df Sum of Sq F Pr(>F)
1 123986 8256.0
2 123985 8225.7 1 30.24 455.8 < 2.2e-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The no regression version, uses proportions tests

```
# A tibble: 3 x 5
group1 group2 p p.adj p.adj.signif
* <chr> <chr> <chr> <chr> <dbl> <dbl> <chr>
1 ctrl email_A 0 0 ****
2 ctrl email_B 0 0 ****
3 email A email B 1e-67 1e-67 ****
```

## Does email A lead to higher average purchases than B?

# A tibble: 2 x 5

	term	estimate	<pre>std.error</pre>	statistic	p.value
	<chr></chr>	<dbl></dbl>	<dbl></dbl>	<dbl></dbl>	<dbl></dbl>
1	(Intercept)	25.6	0.291	88.1	Θ
2	groupemail_B	0.243	0.411	0.592	0.554

## Do the emails lead to higher average purchases?

#	A tibble: 3 >	x 5			
	term	estimate	std.error	statistic	p.value
	<chr></chr>	<dbl></dbl>	<dbl></dbl>	<dbl></dbl>	<dbl></dbl>
1	(Intercept)	12.4	0.268	46.4	Θ
2	groupemail_A	13.2	0.379	34.8	9.47e-265
3	groupemail_B	13.4	0.379	35.5	1.79e-274

## Does email A lead to higher average purchases than B?

Linear hypothesis test

```
Hypothesis:
groupemail_A - groupemail_B = 0
```

```
Model 1: restricted model
Model 2: purch ~ group
```

Res.Df RSS Df Sum of Sq F Pr(>F) 1 123986 367801404 2 123985 367800179 1 1224.8 0.4129 0.5205 The no regression version, uses t-tests

```
# A tibble: 3 x 9
.y. group1 group2 n1 n2 p p.signif p.adj p
* <chr> <chr
```

Email A has significantly higher opens and clicks than email B,

- But purchase are similar for both emails  $\rightarrow$  Send email A!

Both emails generate higher average purchases than the control  $\rightarrow$  Send emails!

## 3/ Design of A/B tests

- 1. Business question
- 2. Test setting (lab vs. field)
- 3. Unit of analysis (visit, customer, store)
- 4. Treatments
- 5. Response variable(s)
- 6. Selection of units
- 7. Assignment to treatments
- 8. Sample size

#### If you can answer these questions, you have a test plan

Business questions: Does email work? If so which email is better?
Test setting: email to retailer customers
Unit: email address
Treatments: email version A, email version B, holdout

**Reponse**: open, click and 30-day purchase (\$)

Selection: all active emails on email list (open in last 12 months)

Assignment: randomly assigned (1/3 each)

Sample size: 123,988 emails

The standard recommendation is to set the sample size **in advance** and not test for significance until the data comes in.

• The recommended sample size is:

$$n_1 = n_2 \approx (z_{1-lpha/2} + z_{eta})^2 \left(rac{2s^2}{d^2}
ight)$$

## Interpreting the sample size formula

$$n_1 = n_2 \approx (z_{1-lpha/2} + z_{eta})^2 \left(rac{2s^2}{d^2}
ight)$$

- More noise,  $s^2 
  ightarrow$  larger sample size
- Smaller difference to detect,  $d \rightarrow$  larger sample size
- Lower error rates,  $(z_{1-\alpha/2} + z_{\beta}) \rightarrow$  larger sample size

**Data is noisy**, so the group with the higher average in the test not always have the higher true response.

There are **two mistakes** you can make:

- **Type I error**: Declare the treatments different, when they are the same ( $\alpha$ )
- **Type II error**: Declare the treatment the same, when they are different  $(\beta)$

I want a low probability of both of those mistakes ( $\alpha$ ,  $\beta$ ) given a specific known difference between treatments (d) and noise in my response (s)

$$n_1 = n_2 \approx (z_{1-\alpha/2} + z_\beta)^2 \left(\frac{2s^2}{d^2}\right)$$

Sample size to detect at \$1 difference in average 30-day purchases:

• Continous response (e.g. money, time on website)

$$n_1 = n_2 \approx (z_{1-\alpha/2} + z_\beta)^2 \left(\frac{2s^2}{d^2}\right)$$

• Binary response (e.g. conversions)

$$n_1 = n_2 \approx (z_{1-\alpha/2} + z_\beta)^2 \left(\frac{2p(1-p)}{d^2}\right)$$

#### **Binary response**

## There are different sample size formulas floating around.

- These formulas differ on what assumptions they may about what you are trying to do,
- It can be very hard to figure out what assumptions are being made
- ... even for experts
- So use some care before plugging numbers into an online calculator

A sample size calculation will help you identify the right amount of data you need for the problem at hand.

## Agreeing on **outcome variables** is **not** as **easy** as it sounds

- Should be defined using short-term metrics that predict long-term value
- (and hard to game)
- Think about customer lifetime value, not immediate revenue
- Use few but key metrics Conversion funnels use Pirate metrics: AARRR: acquisition, activation, retention, revenue, and referral

Experiments at Microsoft (paper):

- 1/3 of ideas were positive ideas and statistically significant
- 1/3 of ideas were flat, with no statistically significant difference
- 1/3 of ideas were negative and statistically significant

At Bing (well optimized), the success rate is lower: 10-20%.

Implication: Aim for small continuous improvements

## Any figure that looks interesting or different is usually wrong

• Check before celebrating

"Experimentation is the least arrogant method of gaining knowledge"

- Isaac Asimov

Some folks believe controlled experiments threaten their jobs

- "we know what to do and we're sure of it"
- Reflex-like rejection of new knowledge because it contradicts entrenched norms, beliefs or paradigm

## Controversy in treatment design

- Facebook's emotional contagion experiment
- Amazon and early pricing experiments
- OK Cupid (Tinder for the previous generation) with deception on match score

## Minimal Risk Experimentation:

"the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests"

When in doubt have an Institutional Review Board

# 4/ Recap

- A/B testing is running Randomized Control Trials online
- Balance tests help confirm that randomization into treatment is indeed random
- Statistical inference toolkit and linear regression enable us to estimate the treatment effects
- The correct sample size for detecting a treatment effect is a crucial aspect of test design
- There are challenges beyond the analysis of data that are important obstacles in implementation

I have borrowed content and inspiration from the following sources:

- Elea Feit's "Advanced A/B testing workshop"
- Ronny Kohavi's "A/B Testing at Scale: Accelerating Software Innovation"

Suggested Citation:

```
@misc{smwa2025_abtest,
    title={"Social Media and Web Analytics: A/B Tests - Basics"},
    author={Lachlan Deer},
    year={2025},
    url = "https://tisem-digital-marketing.github.io/2025-smwa"
}
```

This course adheres to the principles of the Open Science Community of Tilburg University. This initiative advocates for transparency and accessibility in research and teaching to all levels of society and thus creating more accountability and impact.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.