

# A/B Tests: The Essentials

Lachlan Deer

Social Media and Web Analytics, Spring 2024

# Learning Goals

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

# Where Are We Now?

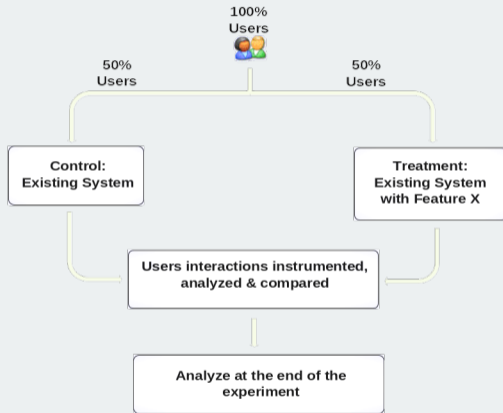
So far we've discussed:

- 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 ↔ Randomized Control Trials online!

- aka Online Controlled Experiments

# A/B Tests: The Basic Idea



Source: Kohavi (2019)

# Example: Bing Ads with Site Links

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

[Esurance® Auto Insurance - You Could Save 28% with Esurance.](#) Ads

[www.esurance.com/California](http://www.esurance.com/California)

Get Your Free Online Quote Today!

A

[Esurance® Auto Insurance - You Could Save 28% with Esurance.](#) Ads

[www.esurance.com/California](http://www.esurance.com/California)

Get Your Free Online Quote Today!

[Get a Quote](#) · [Find Discounts](#) · [An Allstate Company](#) · [Compare Rates](#)

B

**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?

The screenshot shows a Bing search result for "amazon". The search bar contains "amazon" and the results are filtered to "Web". The main result is "Amazon.com - Official Site - Amazon" with a URL that is underlined. Below the main result, there are several sections: "Books", "Kindle eBooks", "Shop All Departments", "Prime Instant Video", "Movies & TV", "Amazon Local", "Electronics", "Shopping", "Full Store Directory", "Outlet", and "News about Amazon". Each section contains a brief description and a link to the respective category. The links are underlined, and the text is blue. The "Stock price" section shows a price of \$427.63, down 4.5% from the previous day. The "Recent post" section shows a post about National Video Day. The "People also search for" section lists eBay, Flipkart, Google, Apple Inc., and Alibaba Group.

The screenshot shows a Bing search result for "amazon". The search bar contains "amazon" and the results are filtered to "Web". The main result is "Amazon.com - Official Site - Amazon" with a URL that is not underlined. Below the main result, there are several sections: "Books", "Kindle eBooks", "Shop All Departments", "Prime Instant Video", "Movies & TV", "Amazon Local", "Electronics", "Shopping", "Full Store Directory", "Outlet", and "News about Amazon". Each section contains a brief description and a link to the respective category. The links are not underlined, and the text is blue. The "Stock price" section shows a price of \$427.63, down 4.5% from the previous day. The "Recent post" section shows a post about National Video Day. The "People also search for" section lists eBay, Flipkart, Google, Apple Inc., and Alibaba Group.

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.

The screenshot shows the Total Wine & More website interface. At the top, there is a navigation bar with a search bar containing the text "What can we help you find today?" and a "Search" button. To the right of the search bar, there is a location dropdown menu set to "West Orange, NJ" and a shopping cart icon. Below the navigation bar, there are several menu items: "Wine", "Spirits", "Beer", "Accessories & More", "Deals", and "Gift Guide". On the right side of this menu, there are links for "Sign In" and "Create Account".

The main content area is titled "Syrah/Shiraz". On the left side, there is a "Stores" filter section with a list of locations and their distances: "West Orange, NJ (0.0 miles)", "Union, NJ (7.6 miles)", "River Edge, NJ (13.9 miles)", "Norwalk, CT (49.7 miles)", and "Milford, CT (69.9 miles)". Below this list are links for "View more", "Select all", and "Deselect all". There is also a "Category" filter section with a dropdown arrow, and "Wine Varietal & Type", "Price Range", "Brand", and "Country/State" filters, all with dropdown arrows.

The main product display area is titled "Available to you (169)" and "All stores (618)". It features a "Show" dropdown menu with options for "Pick up in store", "Ship to NJ", and "Pick up or ship". There is also a "Sort by" dropdown menu set to "Most Popular", a "Show" dropdown menu set to "24", and a "View as" dropdown menu. The product display area contains three wine bottles:

- Molly Dooker Shiraz The Boxer**: 750ml, \$26.97. Buttons: "ADD TO CART", "SAVE TO LIST".
- Yellow Tail Shiraz**: 1.5L, \$11.47. Buttons: "ADD TO CART", "SAVE TO LIST".
- Jem Jar Sweet Shiraz**: 750ml, \$7.46. Buttons: "ADD TO CART", "SAVE TO LIST".

Source: Total Wine & More



# Wine retailer email test

**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", "1901Email", "1901Email", "1901Email", "1901Em~
$ group     <chr> "ctrl", "email_B", "email_A", "email_A", "email_A", "email_~
$ email     <lgl> FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRU~
$ open      <dbl> 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0,~
$ click     <dbl> 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ 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 ( $Y_i$ )

- 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?**

~~Did the treatment affect the response?~~

Was the randomization done correctly?

**How can we check randomization with the data at hand?**

# Randomization checks

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

- Also known as “**Balance tests**”

# Randomization checks: Our data

```
# A tibble: 3 x 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         5.95        188.         71.7         73.6
2 email_A       90.2         5.95        188.         73.5         72.1
3 email_B       89.8         5.94        190.         74.8         71.6
# i 2 more variables: syrah_mean <dbl>, cab_mean <dbl>
```

# Randomization checks

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

```
# note: output omitted
df %>%
  select(group, days_since, visits, past_purchase,
         chard, sav_blanc, syrah, cab) %>%
  st(group = 'group', group.test = TRUE)
```



# Randomization checks

**Randomization seems to check out!**

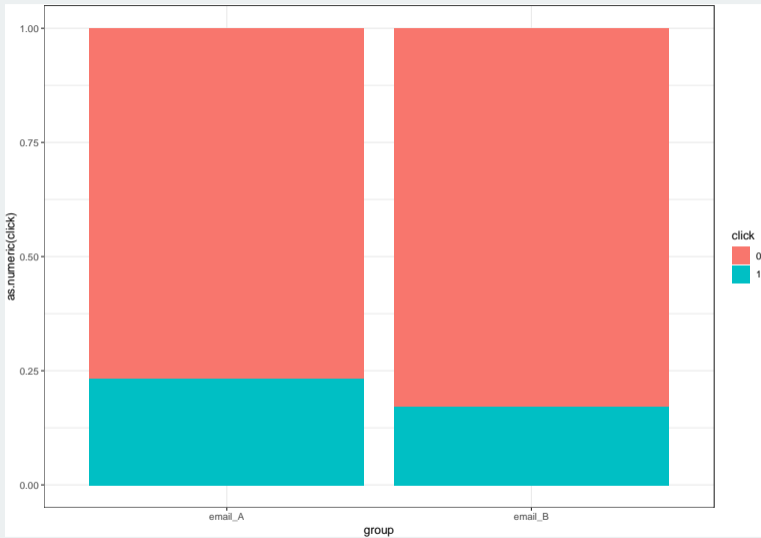
... onto average treatment effects

# Did the treatments affect the responses?

Look at the means of outcome variables between treatments:

```
# A tibble: 3 x 4
  group    open_mean click_mean purch_mean
  <chr>      <dbl>      <dbl>      <dbl>
1 ctrl         0          0          12.4
2 email_A    0.718      0.132      25.6
3 email_B    0.652      0.0934     25.9
```

**Question:** What differences do you observe?



# Does email A have higher open rate than B?

```
# A tibble: 1 x 6
  statistic chisq_df p_value alternative lower_ci upper_ci
  <dbl>     <dbl>   <dbl> <chr>         <dbl>     <dbl>
1     424.         1 1.40e-94 greater      0.0613         1
```

# Does email A have higher open rate than B?

```
# A tibble: 2 x 5
  term          estimate std.error statistic  p.value
<chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    0.718     0.00228     315.    0
2 groupemail_B -0.0666    0.00322    -20.7 1.40e-94
```

# Does email A have higher click rate than B?

```
# A tibble: 2 x 5
  term          estimate std.error statistic  p.value
<chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    0.132     0.00155     84.9  0
2 groupemail_B -0.0383    0.00219    -17.4 6.28e-68
```

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

```
# A tibble: 2 x 5
  term          estimate std.error statistic p.value
<chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    25.6      0.291     88.1     0
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>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	12.4	0.268	46.4	0
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:

$\text{groupemail\_A} - \text{groupemail\_B} = 0$

Model 1: restricted model

Model 2:  $\text{purch} \sim \text{group}$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	123986	367801404				
2	123985	367800179	1	1224.8	0.4129	0.5205

# Summary of findings

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

- But purchase are similar for both emails → Send email A!

Both emails generate higher average purchases than the control → Send emails!

## **3/** Design of A/B tests

# Seven key questions

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**

# Email test

**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

# Sample size planning

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:

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

# Interpreting the sample size formula

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

- More noise,  $s^2 \rightarrow$  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

# Sample size planning: Key ideas

**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 calculator in R

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

```
power.t.test(sd = sd(df$purch), # ideally using
              # pre-experiment data!
              delta = 1, # minimum detectable effect
              sig.level = 0.95, # alpha: industry standard
              power=0.80 # 1 - beta: industry standard
              )
```

# Sample size planning

- **Continuous 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)

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

# Sample size calculator in R

## Binary response

```
power.prop.test(p1=0.07,  
               p2=0.07 + 0.01, # d = 0.01  
               sig.level=0.05,  
               power=0.80  
               )
```

# A word of caution about sample size calculators

There are **different sample size formulas floating around**.

- These formulas differ on what assumptions they may have 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.

# Choosing Outcome Variables

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

# Most Ideas Fail

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**

# Twyman's Law

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

# Summary

- 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

# Acknowledgements

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"

# License & Citation

Suggested Citation:

```
@misc{smwa2024_abtest,  
  title={"Social Media and Web Analytics: A/B Tests - Basics"},  
  author={Lachlan Deer},  
  year={2024},  
  url = "https://tisem-digital-marketing.github.io/2024-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](#).