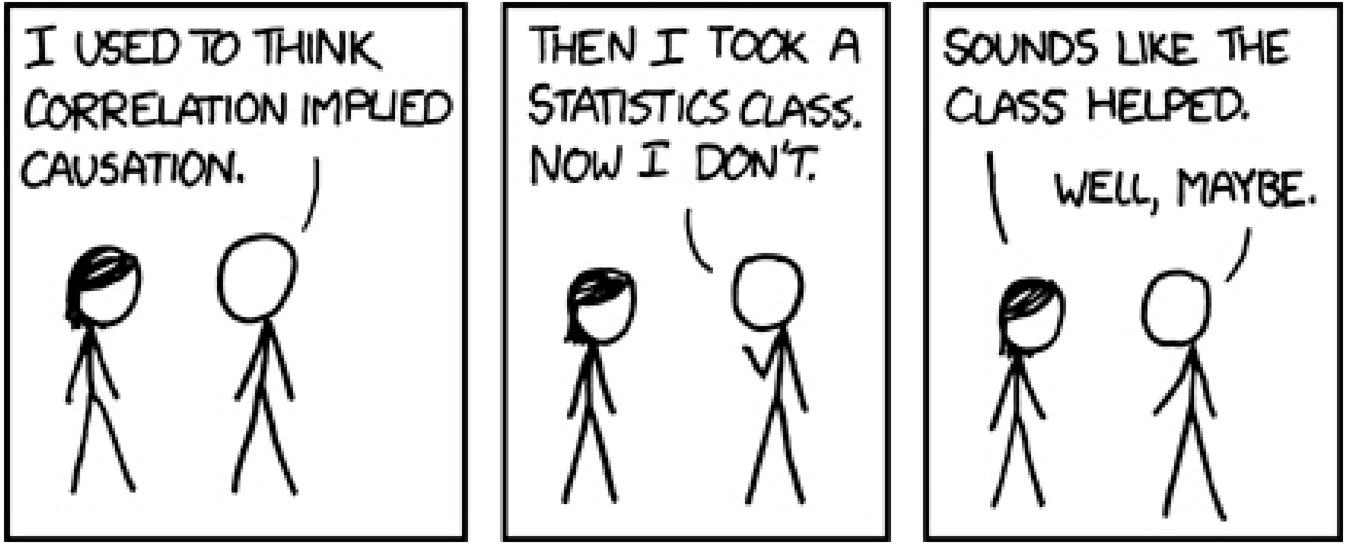
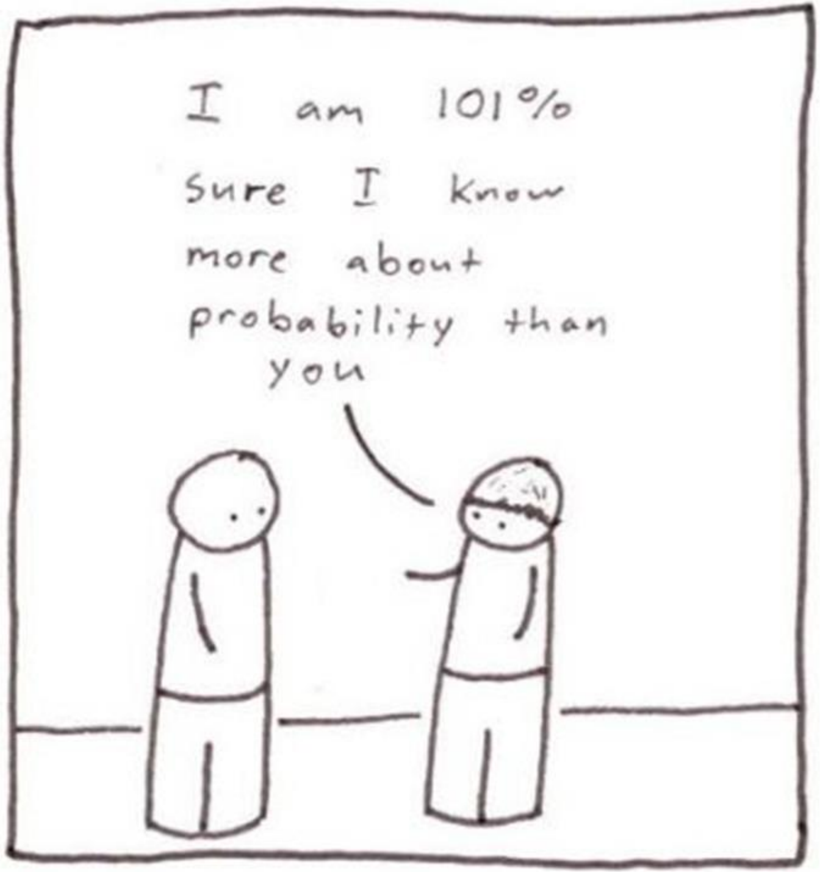
# Introduction to Online Experimentation and A/B Testing

Data Science Dojo

datasciencedojo
unleash the data scientist in you


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.


I am 101% sure I know more about probability than you

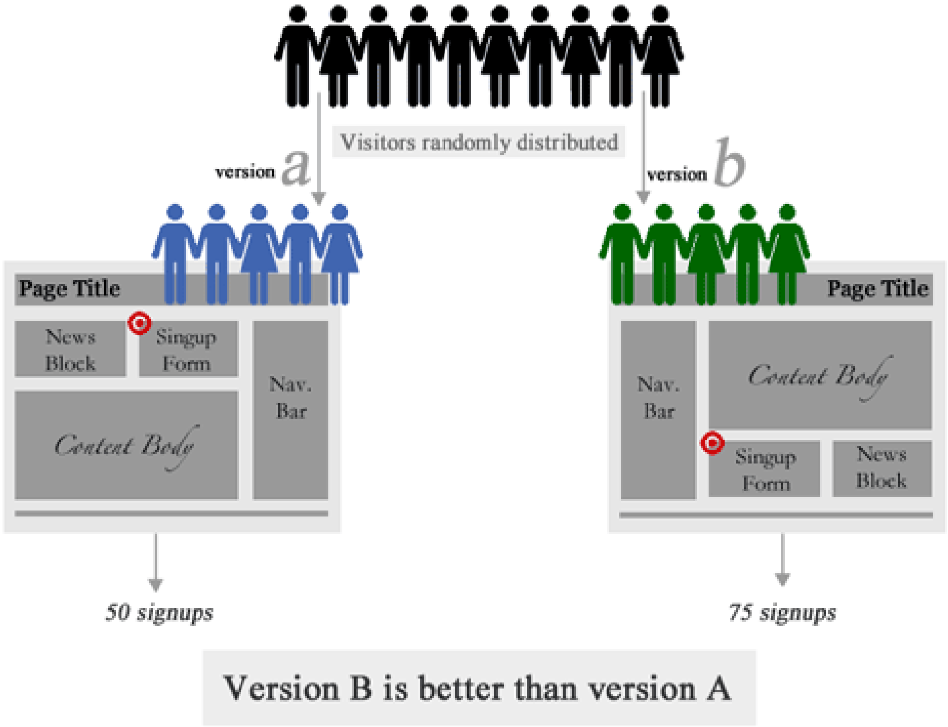
# Agenda

- Introduction
  - What is A/B testing?
  - Some interesting A/B tests
- Fundamentals
  - Hypothesis testing and related ideas
  - Metrics for A/B testing
  - Focus on intuitive understanding than specific distributions, formulas and tests
- Common pitfalls
  - Depth of discussion will depend upon audience engagement and time

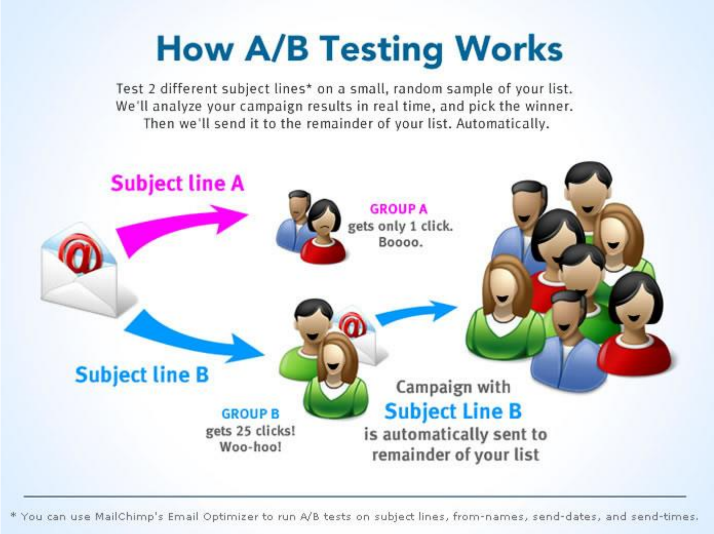
In God we trust. All others bring data.

W. E. Deming

# What is A/B Testing?



# A/B Testing On Newsletters And Email

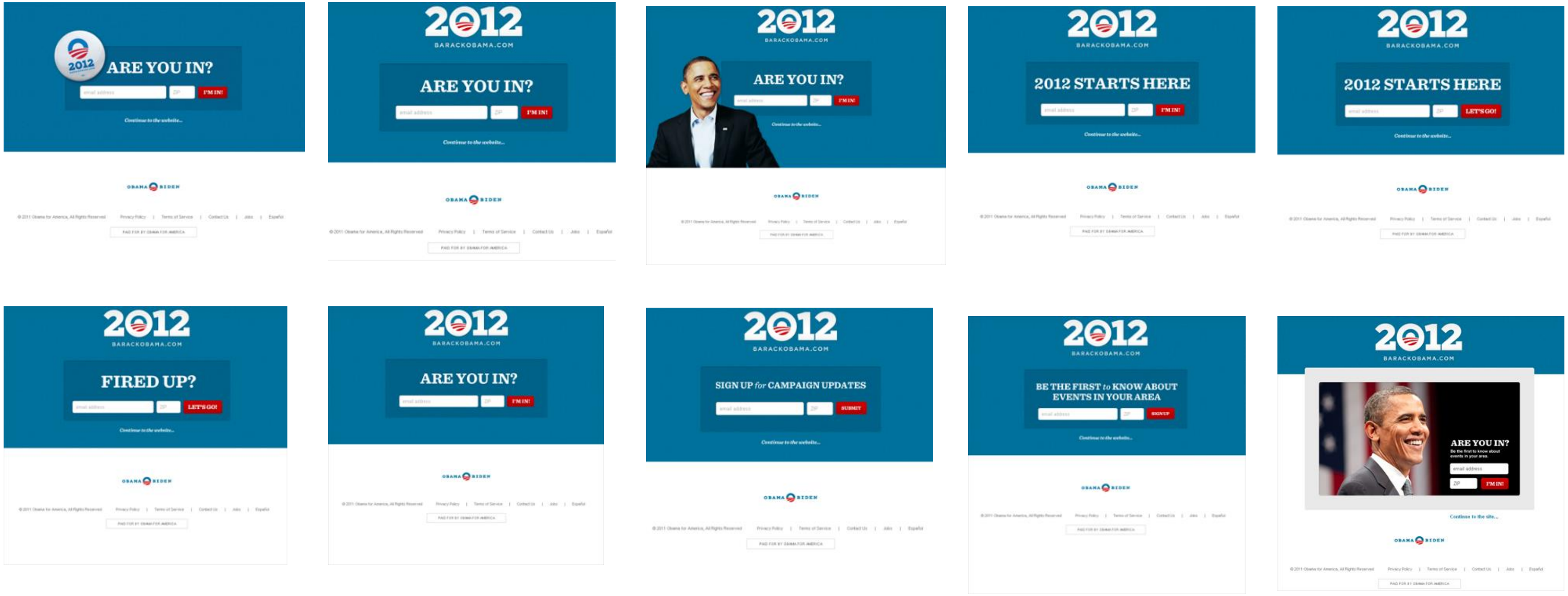



Run tests on many things

- Subject lines
- From names
- Send dates
- Send time

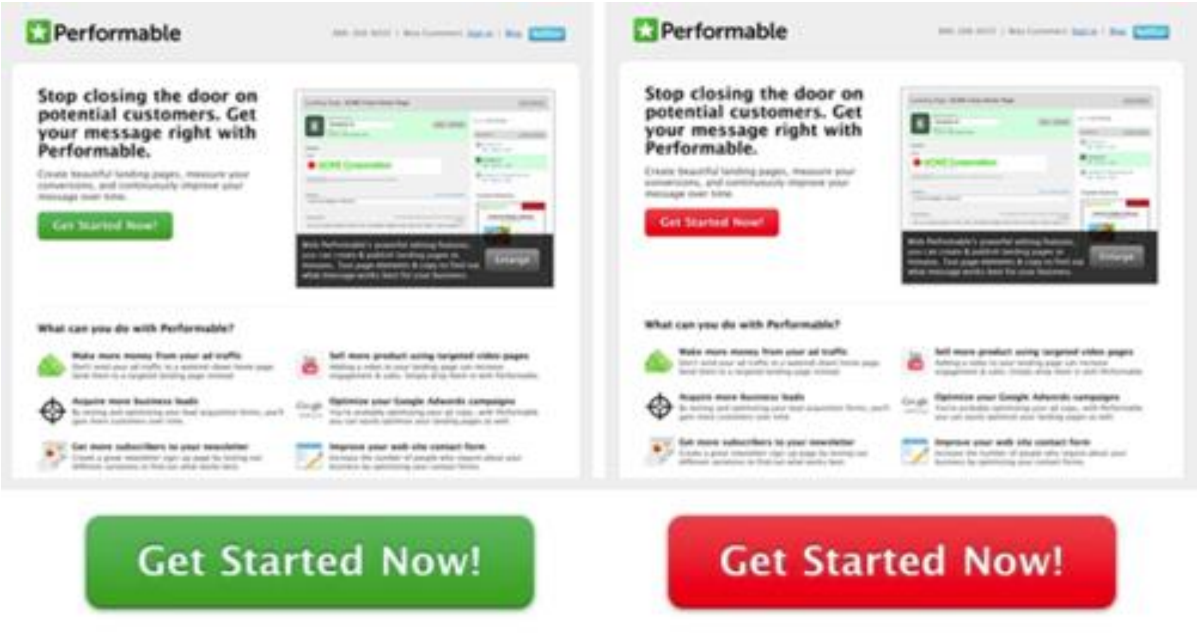
# Obama 2012 Campaign

# Maximize Sign-Ups And Donations




**Source:** http://www.nathanielward.net/2011/06/see-ab-testing-in-action-on-barack-obamas-reelection-website/

# Testing Call-to-Action Button




Red button increased clicks by 21%

# Testing Navigation Bar



'How It Works' increased clicks by 47.7%

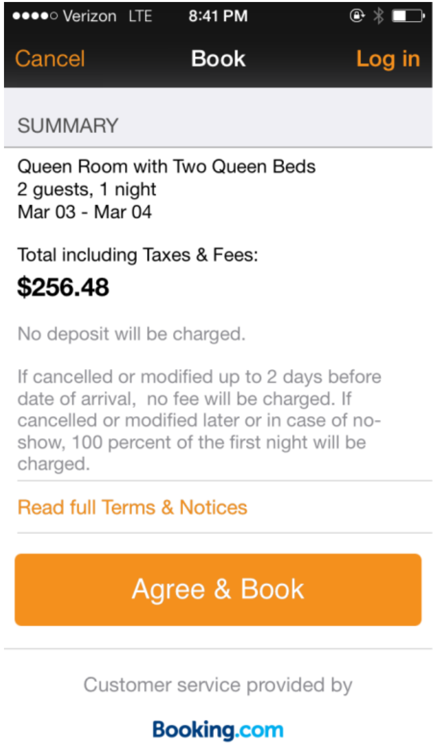
# Jocelyn or Michael

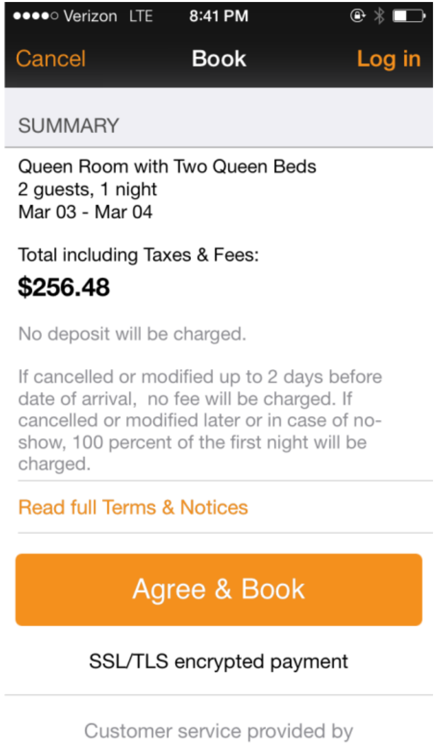

JOCELYN


MICHAEL

Michael increased conversions by 21%

# Kayak: Is reassurance good or bad?




datasciencedojo
unleash the data scientist in you

# Why We Use A/B Testing

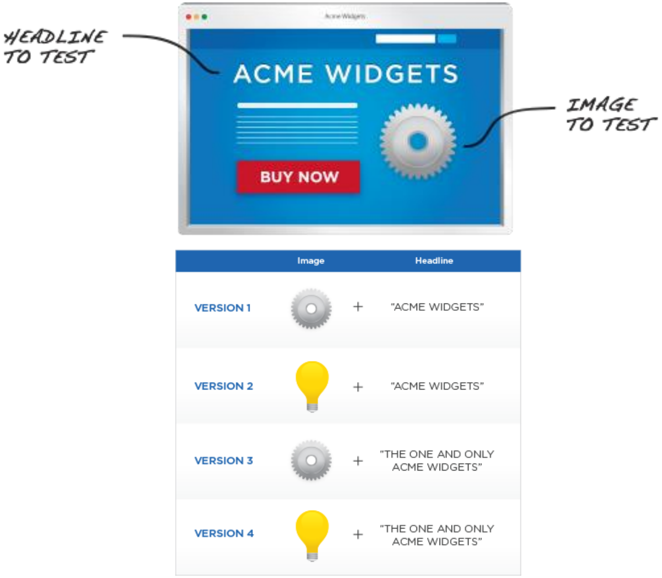
**Human Psychology**

- Users are complex and our intuition is often wrong
- Know what the users want subconsciously or otherwise.
- Impact is always expected to be positive, but outcome is often humbling

**Business**

- Rolling out a feature to all the users at the same time is risky
- Helps fail fast and move on

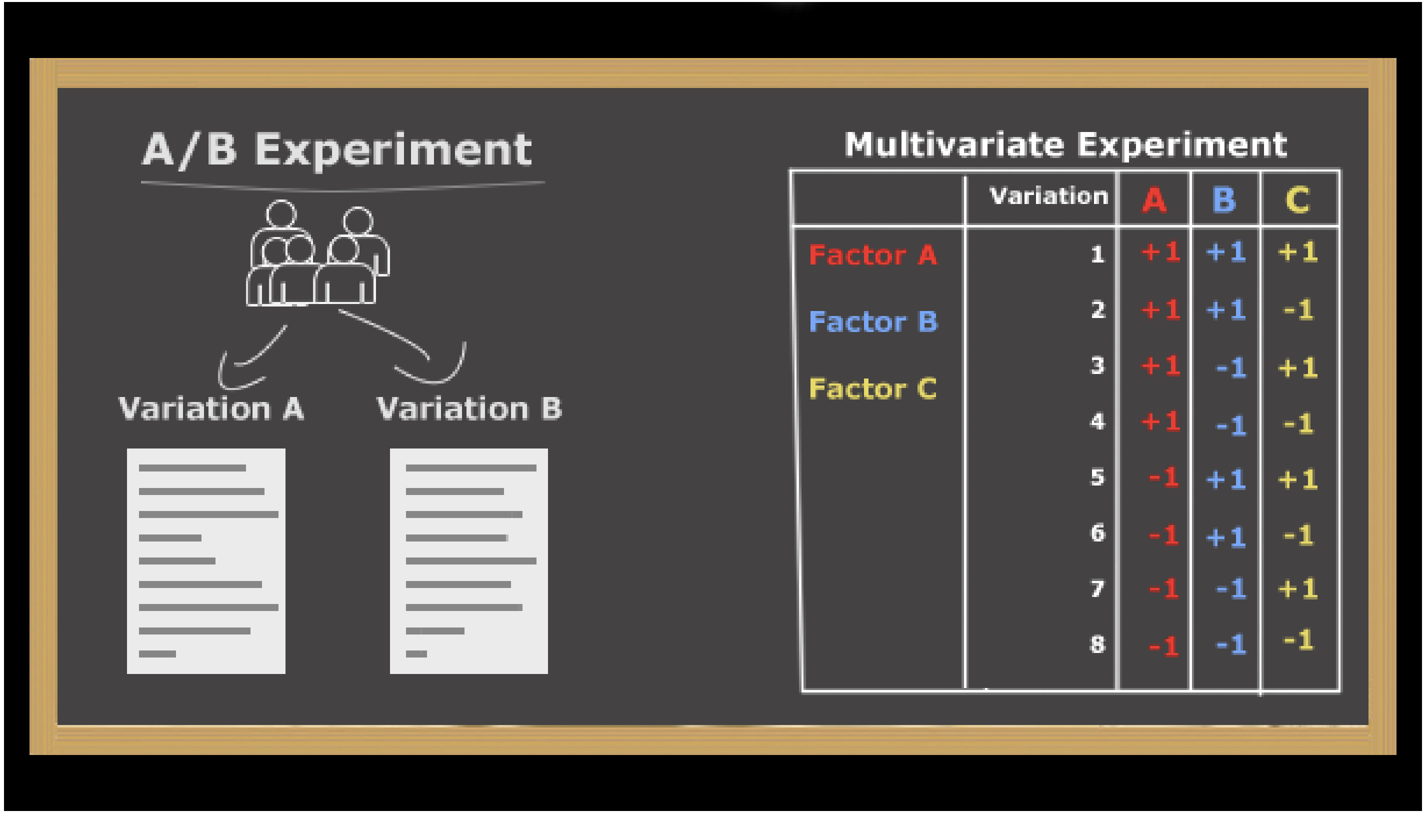
# A/B Testing vs. Multivariate Testing


datasciencedojo
unleash the data scientist in you

# A/B Testing vs Multivariate Testing

| | A/B Testing | Multi-variate Testing |
|---|---|---|
| Common use | Compare two very different designs with each other | Several minor variations are up for debate. Two colors of button with three different headlines.<br>Also called full factorial testing |
| Advantages | Simple in design.<br>Small sample size may be ok. | A lot of different combinations tried at once. |
| Limitations | Trying only one alternative | Bigger sample size.<br>Complex.<br>Need better understanding of interactions |



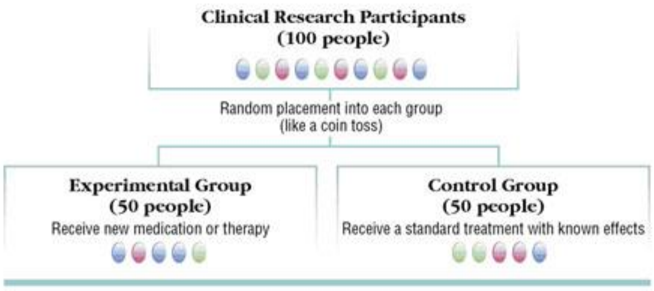

| | Variation | A | B | C |
|---|---|---|---|---|
| Factor A | 1 | +1 | +1 | +1 |
| Factor B | 2 | +1 | +1 | -1 |
| Factor C | 3 | +1 | -1 | +1 |
| | 4 | +1 | -1 | -1 |
| | 5 | -1 | +1 | +1 |
| | 6 | -1 | +1 | -1 |
| | 7 | -1 | -1 | +1 |
| | 8 | -1 | -1 | -1 |

# Terminology

# Control and Treatment

- Control
  - Default experience. The way things are now.
  - Example: Current look and feel of your ‘Buy Now’ button


- Treatment
  - The change we want to make
  - Example: Change the button from green to blue

# Factor and Level

- Factor
  - The item we want change

- Level
  - The variations of factor


datasciencedojo
unleash the data scientist in you

# Metrics Used For A/B Testing

- Search engines:
  - Queries/UU, Session length, Sessions/UU, Page views, Bounce rate
- Online Retailers:
  - Conversion rate, revenue/UU, Avg Cart Value and so on
- Other websites:
  - CTR, signup for newsletter
- Each business is different

# Brainstorm

# OEC: Overall evaluation Criteria

- Summarizes the primary indicator of success
- May be one of the metrics or a combination of metrics

# Null vs Alternate Hypothesis

- Null Hypothesis ($H_o$):
  - Control and treatment are similar (in terms of the parameter we are estimating)
- Alternate Hypothesis ($H_a$):
  - Treatment is different from control

# Null vs Alternate Hypothesis


Control    Treatment

- Null Hypothesis ($H_o$): Green and blue buttons have the same CTR
- Alternate Hypothesis ($H_a$): CTRs for both buttons are different

# Type I and Type II Error

**Type I Error**
The probability of **falsely accepting** null hypothesis

**Type II Error**
The probability of **falsely rejecting** null hypothesis

| | | Ground Truth | |
|---|---|---|---|
| | | Ho is true. | Ho is false. |
| Experiment Outcome | Reject Ho. | Type I error | Correct decision. |
| | Do not reject Ho. | Correct decision. | Type II error |

# Power

- Power of an online experiment is the probability of **not** rejecting the null hypothesis false
- Which is really 1 – Probability (Type II Error)

# Can you tell me in simple words...

# The Cook and Smoke Detector

- Null Hypothesis (Ho): There is no fire
- Alternate Hypothesis (Ha): There is fire

datasciencedojo
unleash the data scientist in you

# The Cook and Smoke Detector

- Type I Error: There is no fire but smoke detector goes off.
- The cook removes the alarm to prevent type I error.
- This increases the chance of Type II Error i.e. a fire without an alarm.

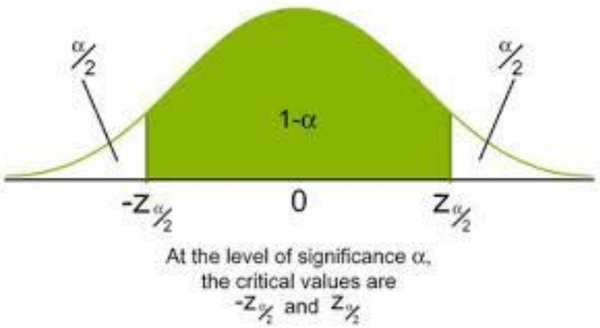
# The Boy Who Cried Wolf

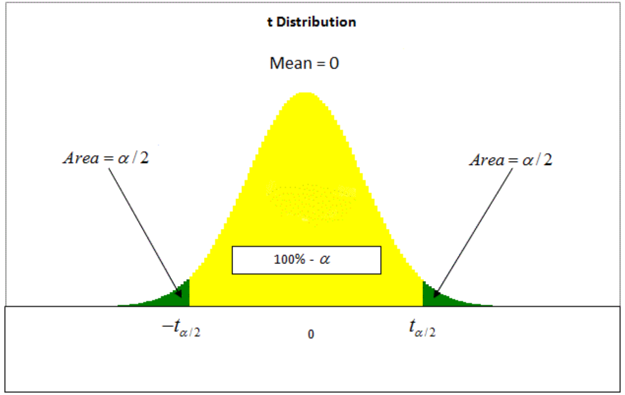
- Null Hypothesis (Ho): There is no wolf
- Alternate Hypothesis (Ha): There is wolf

# The Boy Who Cried Wolf

- Type I Error: Villagers believing the boy when there is no wolf
- Type II Error: Villagers not believing the boy when wolf is there

# Calculating Confidence Interval




| Confidence level | Z score |
|---|---|
| 90% | 1.645 |
| 95% | 1.960 |
| 98% | 2.326 |
| 99% | 2.576 |

**Critical Values (t*)**

| | Confidence Level | | |
|---|---|---|---|
| n − 1 | 0.900 | 0.950 | 0.990 |
| 10 | 1.812 | 2.228 | 3.169 |
| 20 | 1.725 | 2.086 | 2.845 |
| 30 | 1.697 | 2.042 | 2.750 |
| 40 | 1.684 | 2.021 | 2.704 |
| 50 | 1.676 | 2.009 | 2.678 |
| 60 | 1.671 | 2.000 | 2.660 |
| 70 | 1.667 | 1.994 | 2.648 |
| 80 | 1.664 | 1.990 | 2.639 |
| 90 | 1.662 | 1.987 | 2.632 |
| 100 | 1.660 | 1.984 | 2.626 |

# Type I and Type II Error

**Type I Error**

The probability of **falsely accepting** null hypothesis

**Type II Error**

The probability of **falsely rejecting** null hypothesis

Ground Truth

Experiment Outcome

| | Ho is true. | Ho is false. |
|---|---|---|
| Reject Ho. | Type I error | Correct decision. |
| Do not reject Ho. | Correct decision. | Type II error |

# Type I and Type II Error

**Type I Error**

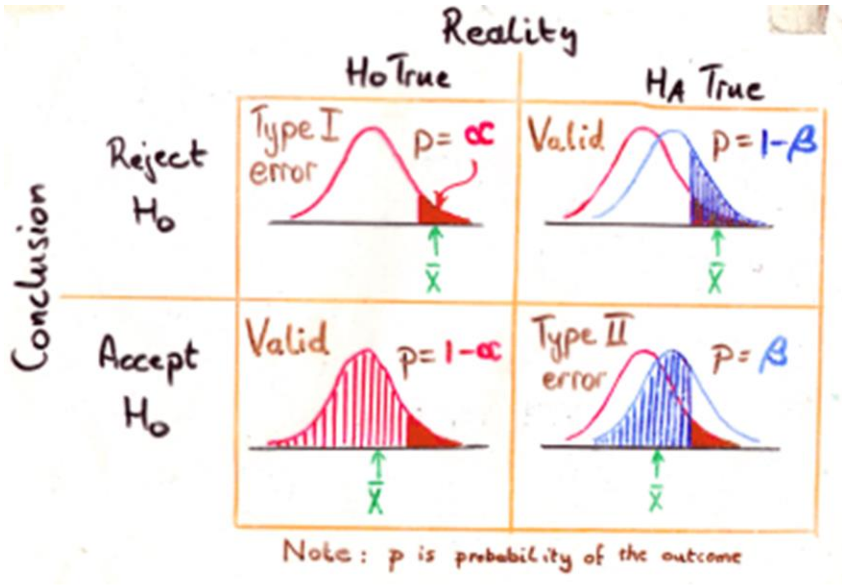
The probability of **falsely accepting** null hypothesis

**Type II Error**

The probability of **falsely rejecting** null hypothesis

Ground Truth

Experiment Outcome



# Confidence Interval

- Range of plausible values of parameter being estimated given the sample data


datasciencedojo
unleash the data scientist in you

# A/A Test

- Comparing the identical experience on different random set of users
- Used for validation of setup

Control

Treatment

# Steps in Experimentation

## Planning
- Choose factors, levels, sample size(how long to run)
- What business question to answer
- Metrics and expected outcome
- Who is in experiment?

## Coding and Logging
- Setup of test and instrumentation

## A/A Test
- To make sure the setup is correct.

## A/B and/or multivariate test

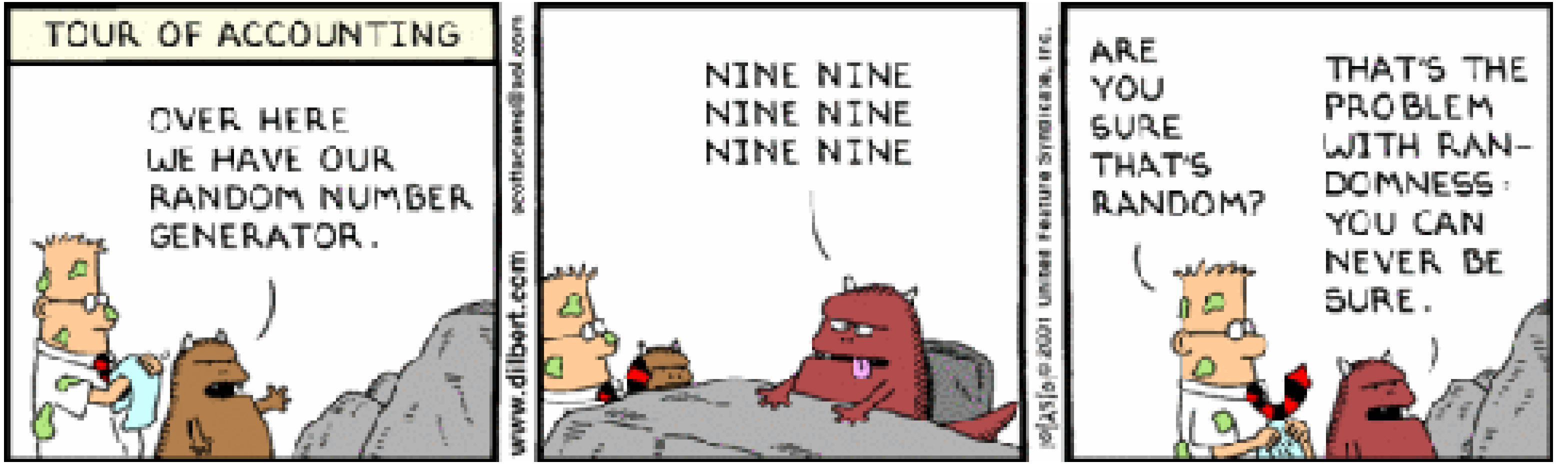
## Analysis and interpretation
- Some times this can be an art
- Newness effect
- Seasonality, segments etc.

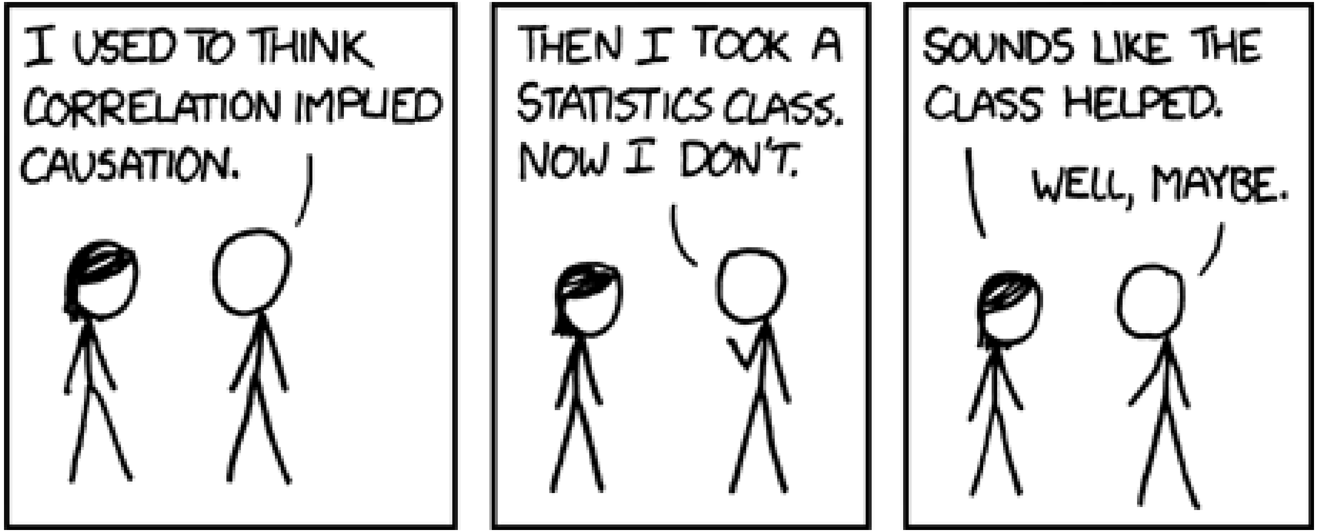
## Make a Decision
- To ship or not to ship

# Categories of Metrics

| | Short-term | Medium-term | Long-term |
|---|---|---|---|
| Examples | CTR<br>PVs<br>Bounce Rate | PVs/user/day<br>CTR/user /day<br>Avg session length | Days with at least one visit,<br>Total time on site<br>Repeat visits/user |
| What is measured? | Immediate or almost immediate impact | Engagement over hours up to a day | Loyalty |

datasciencedojo
unleash the data scientist in you


TOUR OF ACCOUNTING
OVER HERE WE HAVE OUR RANDOM NUMBER GENERATOR.
NINE NINE NINE NINE NINE NINE
ARE YOU SURE THAT'S RANDOM?
THAT'S THE PROBLEM WITH RAN-DOMNESS: YOU CAN NEVER BE SURE.
www.dilbert.com




datasciencedojo
unleash the data scientist in you

# Pitfalls in Online Experimentation

- **Pitfall 1:** Picking an OEC for which it is easy to beat the control
- **Pitfall 2:** Incorrectly computing the confidence intervals
- **Pitfall 3:** Using standard statistical formulas for computation of variance and power
- **Pitfall 4:** Combining metrics over periods where proportions assigned to Control and Treatment vary or over subpopulations sampled at different rates
- **Pitfall 5:** Neglecting to filter bots
- **Pitfall 6:** Failing to validate each step of the analysis pipeline and the OEC components
- **Pitfall 7:** Forgetting to control for all differences, and assuming that humans can keep the variants in sync

# Pitfall 1: Picking an Easy-to-Beat OEC

- Before running an experiment an OEC is selected
- OEC should be tied to a long term goals as opposed to short term goals. CTR vs. long term revenue
- Loyal/repeat users get more weight?
- Sometimes getting the true metric is hard. High CTR does not necessarily mean high conversion rate

# Pitfall 1: Picking an Easy-to-Beat OEC

- Measuring click through on a small area of the page, ignoring the impact on other areas
  - What if the small area on the page was bold/flashing/high contrast?
  - What happens to the whole page CTR?
- Is 'time on site' a good OEC?
  - What if the treatment has a reduced user's effectiveness?

# Pitfall 2: Incorrect Computation of Confidence Intervals

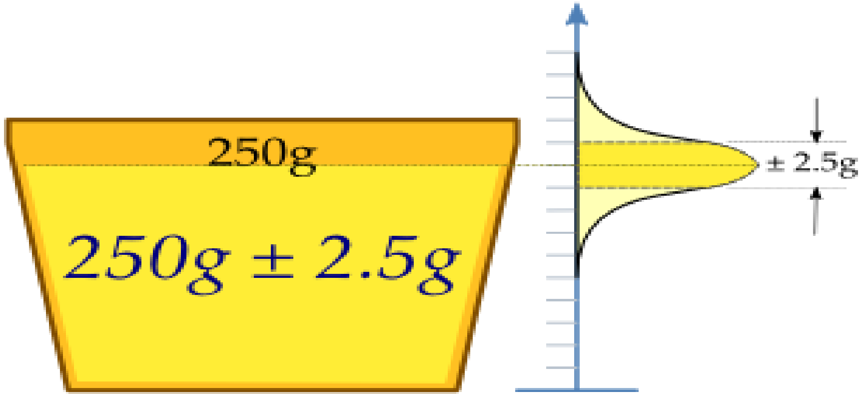
- Hypothesis Test: determines whether there is a statistically significant difference in the means of the control and the treatment
- Confidence Interval: provides a plausible range of the size of the effect (difference in C and T means)

# Pitfall 2: Incorrect Computation of Confidence Intervals

$$0.95 = 1 - \alpha = P(-z \le Z \le z) = P\left(-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right)$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{X} - 1.96 \times 0.5 \le \mu \le \bar{X} + 1.96 \times 0.5\right)$$




$$(\bar{x} - 0.98; \bar{x} + 0.98) = (250.2 - 0.98; 250.2 + 0.98) = (249.22; 251.18).$$

**Confidence interval implies:** If we randomly fill a cup from this vending machine, there is a 95% chance that our cup will have this much coffee

datasciencedojo
unleash the data scientist in you

# Pitfall 2: Incorrect Computation of Confidence Intervals

- Confidence interval should be formed out of absolute difference
- Do not form a confidence interval around percent change. Percentage change involves dividing by a random variable.
- Some techniques to compute CI are mentioned when the OEC is a linear/non-linear combination of metrics that have the same/different basis/experimental unit.

# Pitfall 3: Standard Statistical Formulas for Computation of Variance and Power

- Variance of the metric is needed to compute the statistical significance
- Variance estimates using standard statistical formula for some families of metrics are inaccurate
- This happens when the experimental unit used in random assignment is different from the experiment unit used in the calculation of the metric.

# Pitfall 3: Standard Statistical Formulas for Computation of Variance and Power

- Variance, Power and Sample size estimates may be wrong if care is not taken
- How to correct this?
  - Bootstrap method: Estimate variance using bootstrap samples and compare with the variance from standard formula
- This should be done for all metrics and especially for the one with different experiment and randomization units

# Pitfall 4: Simpson's Paradox

- Unintuitive but not uncommon
- Simpson's paradox: 'A correlation or trend present in different groups is reversed when the groups are combined'.

| | Treatment A | Treatment B |
|---|---|---|
| Small Stones | Group 1 | Group 2 |
| | 93% (81/87) | 87% (234/270) |
| Large Stones | Group 3 | Group 4 |
| | 73% (192/263) | 69% (55/80) |
| Both | 78% (273/350) | 83% (289/350) |

datasciencedojo
unleash the data scientist in you

# Pitfall 4: Simpson's Paradox

- 1 million visitors/day
- On Friday the treatment ran with 1% traffic
- On Saturday, the allocation was raised to 50%.
- If we consider Friday and Saturday separately T has a better CTR
- T's CTR is worse when aggregated over days

**Table 1: Conversion Rate for two days.**
**Each day has 1M customers, and the Treatment (T) is better than Control (C) on each day, yet worse overall**

| | Friday<br>C/T split: 99%/1% | Saturday<br>C/T split: 50%/50% | Total |
|---|---|---|---|
| C | $\frac{20{,}000}{990{,}000} = 2.02\%$ | $\frac{5{,}000}{500{,}000} = 1.00\%$ | $\frac{25{,}000}{1{,}490{,}000} = 1.68\%$ |
| T | $\frac{230}{10{,}000} = 2.30\%$ | $\frac{6{,}000}{500{,}000} = 1.20\%$ | $\frac{6{,}230}{510{,}000} = 1.20\%$ |

It is possible to have $\frac{a}{b} < \frac{A}{B}$ and $\frac{c}{d} < \frac{C}{D}$ while $\frac{a+c}{b+C} < \frac{A+C}{B+D}$

# Pitfall 4: Simpson's Paradox – A Scenario in Controlled Experiments

- Sampling of users with non uniform sampling to make sure all browsers have a representative sample
- Overall results show treatment is better than control but when segmented by browser, control looks better than treatment for each browser

# Pitfall 5: Ignoring Bot Traffic

- For experimentation, we are interested in removing bots/fraud clicks that are not uniformly distributed across the control and treatment
- Uniformly distributed bots will only reduce the power of the experiment

# Pitfall 5: Ignoring Bot Traffic

- Failing to exclude bot traffic and fraud clicks may invalidate the results of an experiment

# Pitfall 6: Failing to Validate Each Step of Analysis

- It is important to keep a check on the health of the pipeline
  - Assignment of users to experiment variants
  - Calculation of metrics
  - Any abnormal shift in metrics
  - Movement of metrics that are not expected to move
  - Broken instrumentation

# Pitfall 6: Failing to Validate Each Step of Analysis

- Logging Tests:
  - Compare with real historical data
  - Compare with generated data
  - Look for unexpected patterns
    - Volume of data over time
    - New and repeat users over time
    - Abnormal shift in any of the metrics
  - A/A Tests
  - Rich Instrumentation

# Pitfall 7: Failing to ‘Control’ the Control

- Don’t allow any difference between the Control and the treatment besides what is actually being tested
- If the treatment has some updates, control should have them too and vice versa

# Pitfall 7: Failing to ‘Control’ the Control

- If the site is receiving frequent updates, these updates should be applied equally to the control and the treatment
- Forgetting to control for all differences, and assuming that humans can keep the variants in sync.

# A/B Testing Tools

Google Analytics

Optimizely

unbounce

Visual Website Optimizer

KISSmetrics

Have you heard the latest statistics joke?

Probably....

Did you hear about the statistician who was thrown in jail?

He now has zero degrees of freedom.

datasciencedojo
unleash the data scientist in you

A statistician's wife has twins. He was delighted, and he called to tell his minister the good news.

"Excellent!", said the minister. "Bring them to church on Sunday and we'll baptize them.“

"No," replied the statistician. "Let's just baptize one. We'll keep the other as control."

datasciencedojo
unleash the data scientist in you

How many statisticians does it take to change a light bulb?

**1 – 3.  $\alpha=0.05$**

# Questions?


datasciencedojo
unleash the data scientist in you