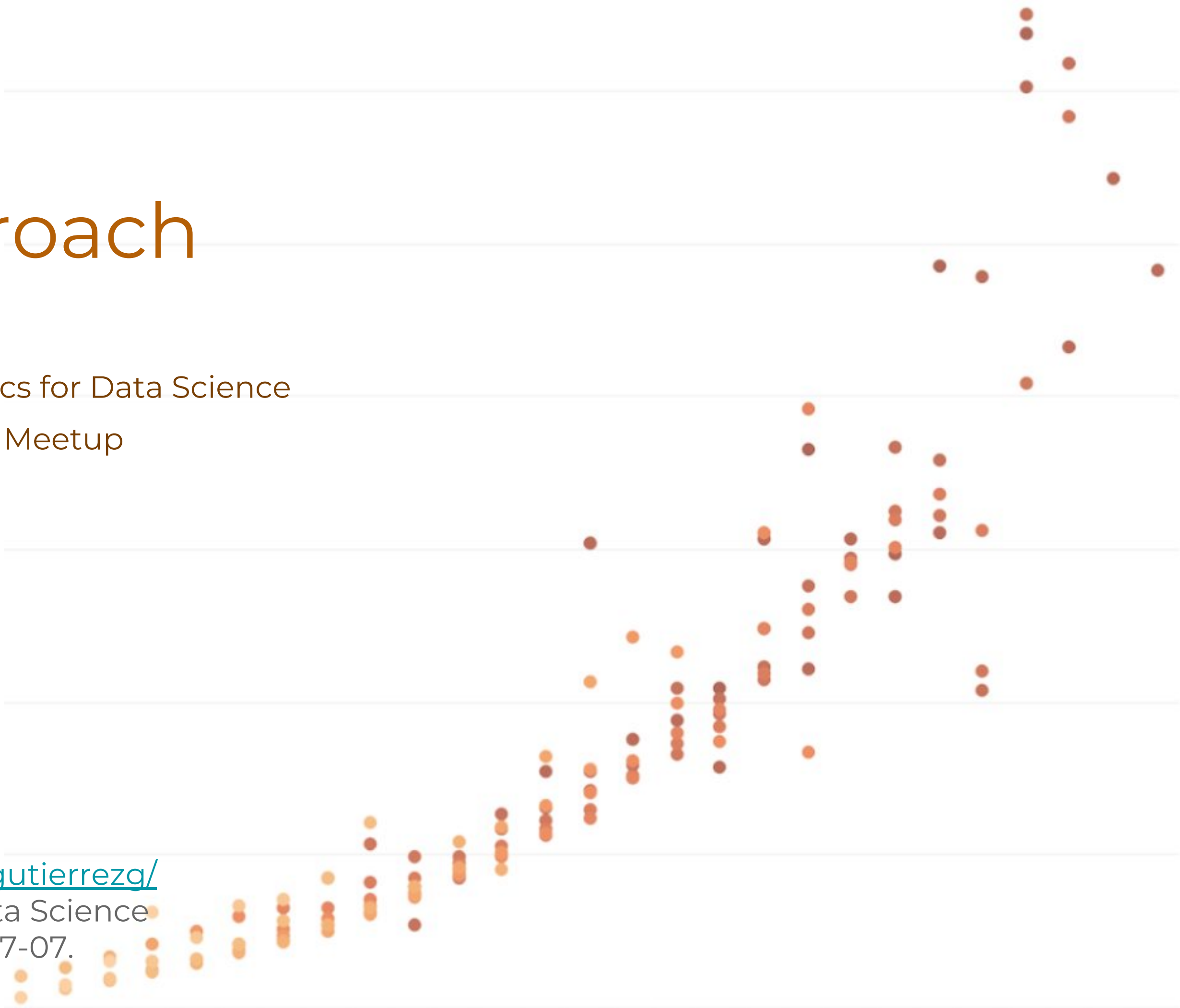
Statistical Approach for A/B Testing

Applied Math, Probability and Statistics for Data Science

Data Science and Engineering Club - Meetup

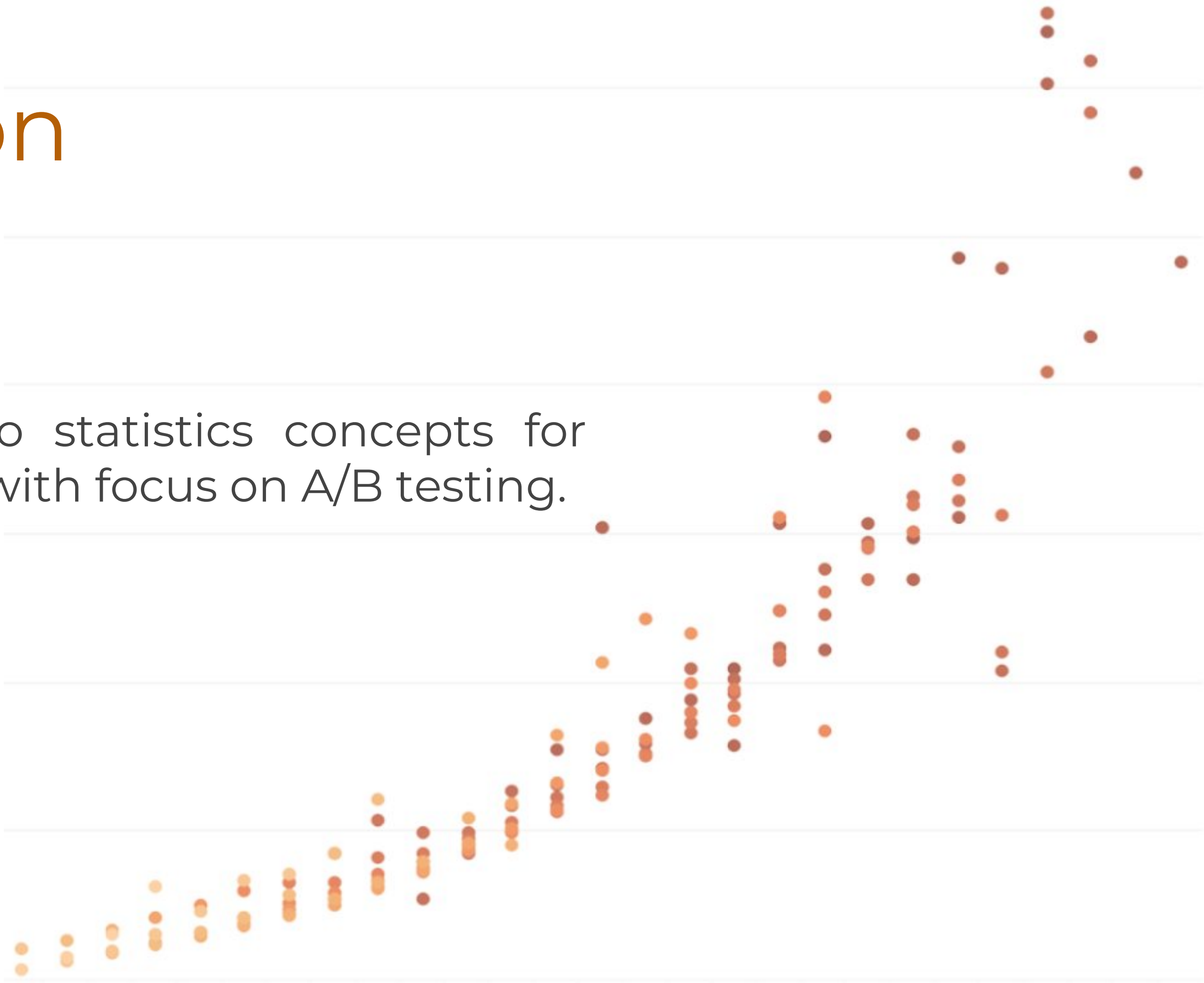


Daniela Gutierrez
dani.gutierrez.g@gmail.com
https://www.linkedin.com/in/danielagutierrezg/
Optics, Physics(MS) Analytics and Data Science

Bank of Ireland Trinity branch, 2018-07-07.

Introduction

A general approach to statistics concepts for data experimentation, with focus on A/B testing.



You are testing real People

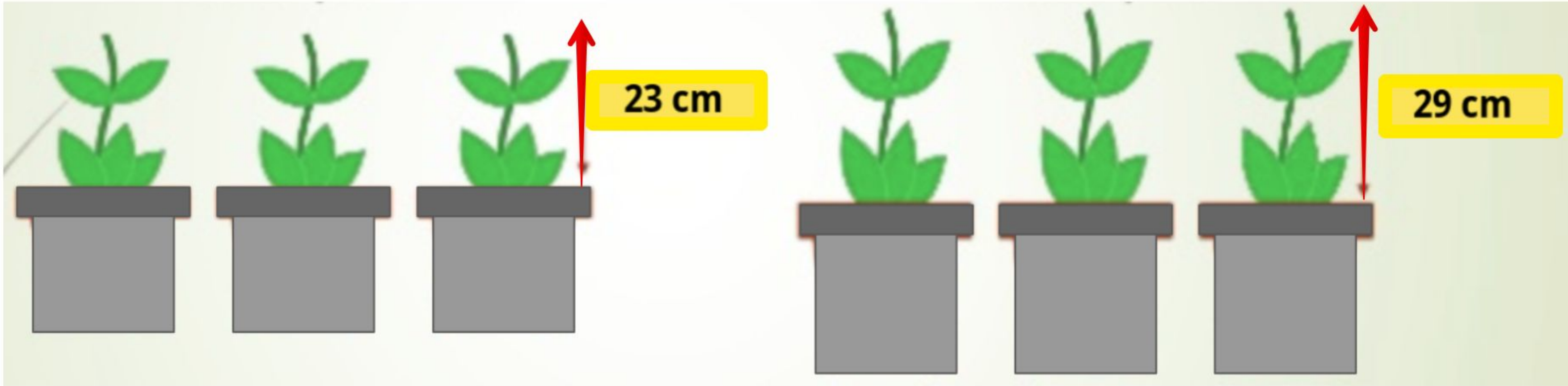
Consumer, Users, ID’s, UUID’s, Customers … is real people!

People has so many variables we can’t control, this make human behaviour very difficult for measurements.

What is an A/B Test?

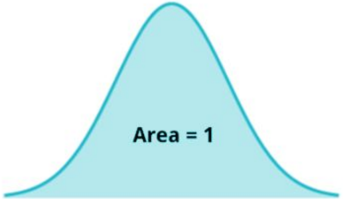
Control Group
'Normal', as always
Used to compare results

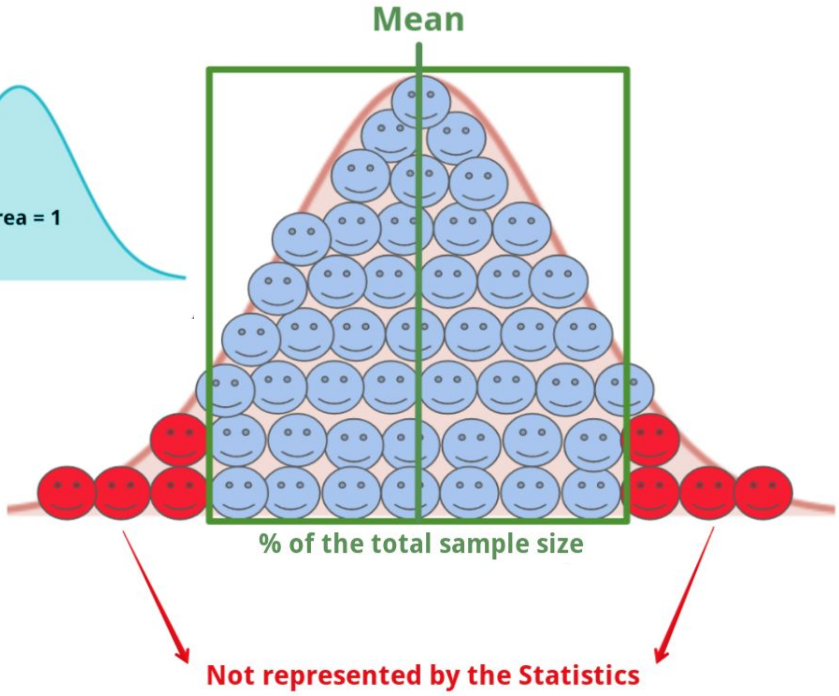
Treatment Group
'Tested', with the variation
Used to measure any difference



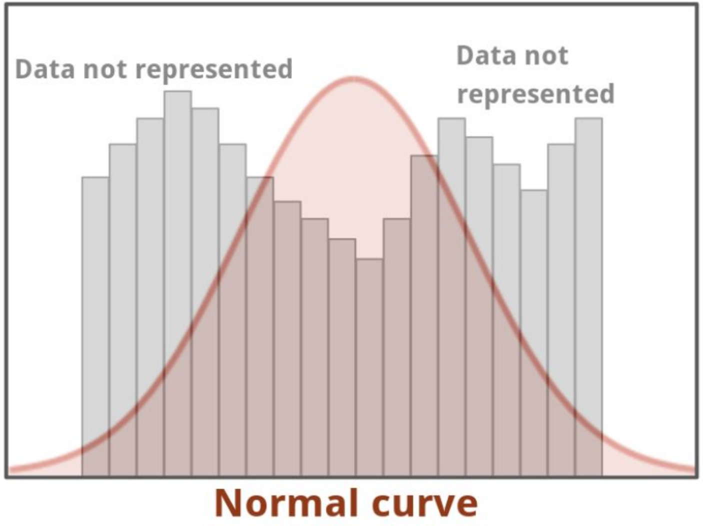
Normal and Non-Normal Distributions

Normal Distribution





Not Normal Distribution



Non-Normal Distributions

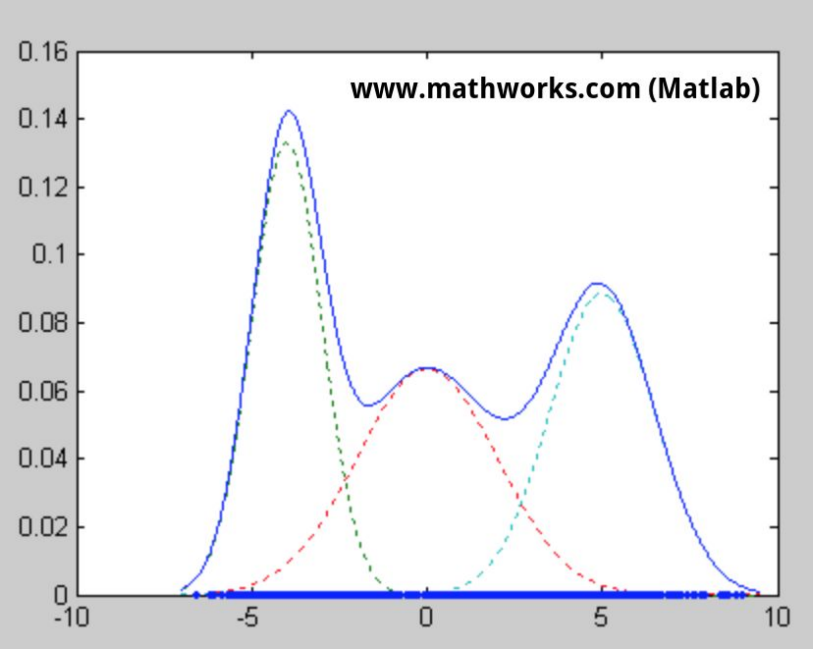
Non-normal data is not very common.
However, if you have it could be because:

Extreme values (many outliers)

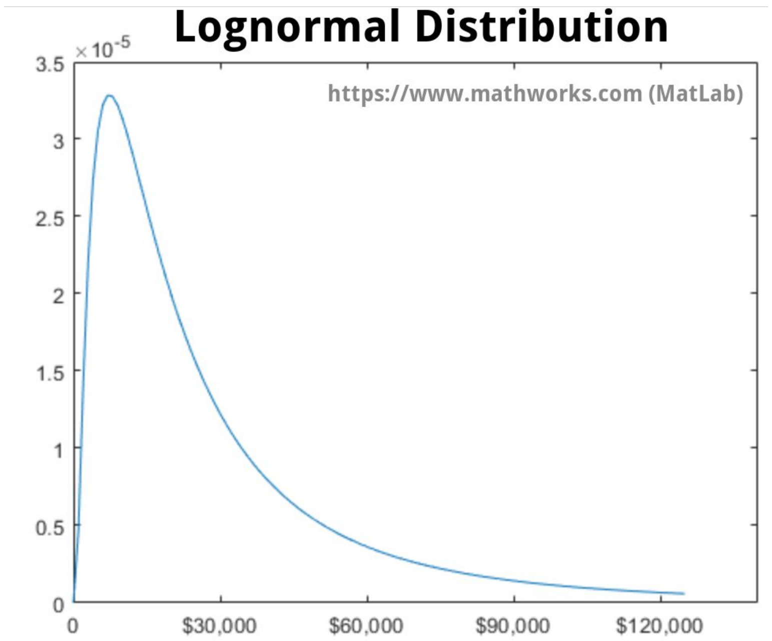
Overlapping because your data has different sources/process or insufficient data discrimination.

Values are close to a natural limits (or zero)

Different distributions (Lognormal, Exponential, Binomial).

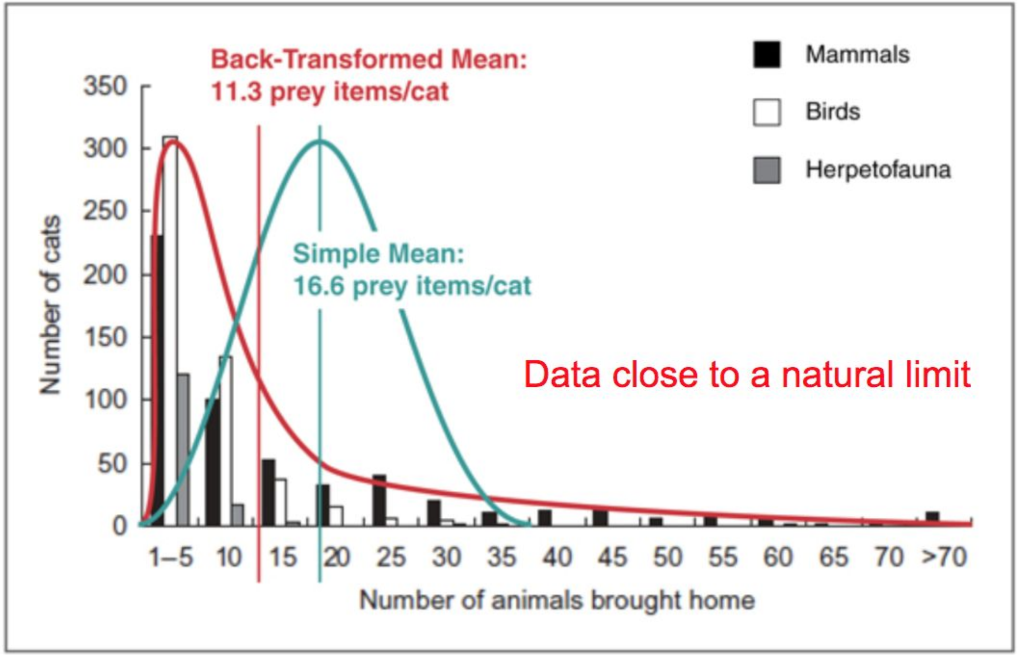


Examples



Lognormal Distribution

https://www.mathworks.com (MatLab)



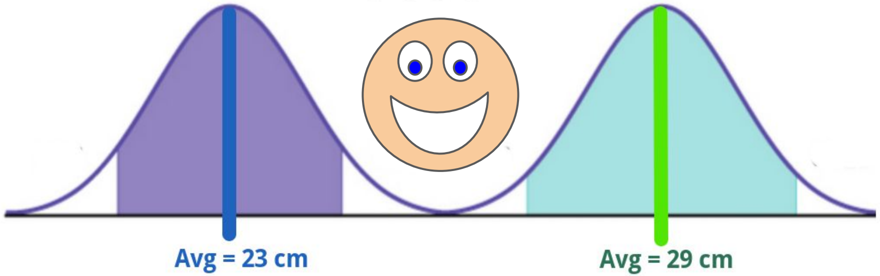
Data close to a natural limit

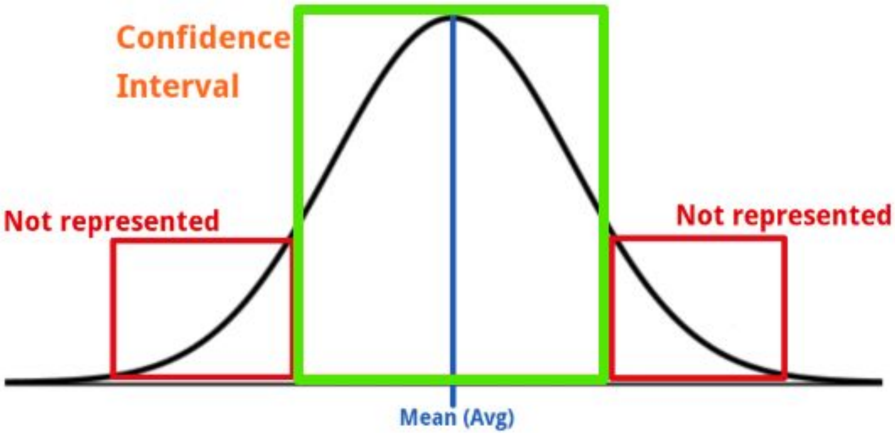
Adapted from Woods et al. 2003

source www.voxfelina.com

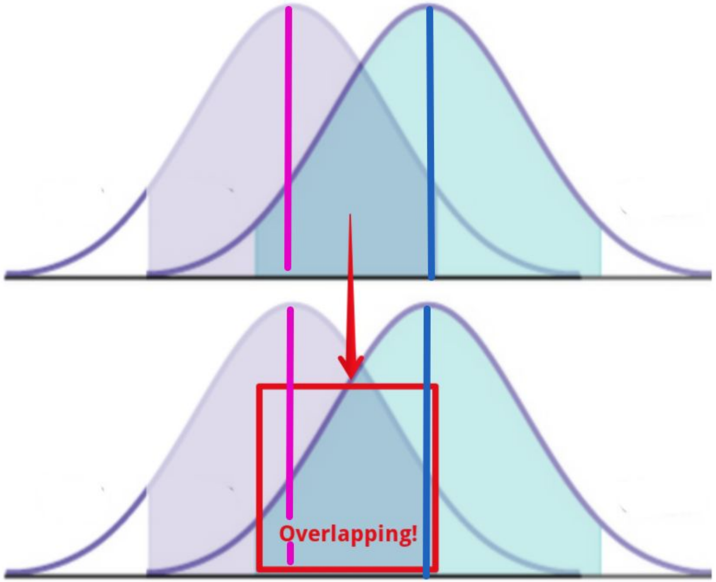
Confidence Interval

Case 1



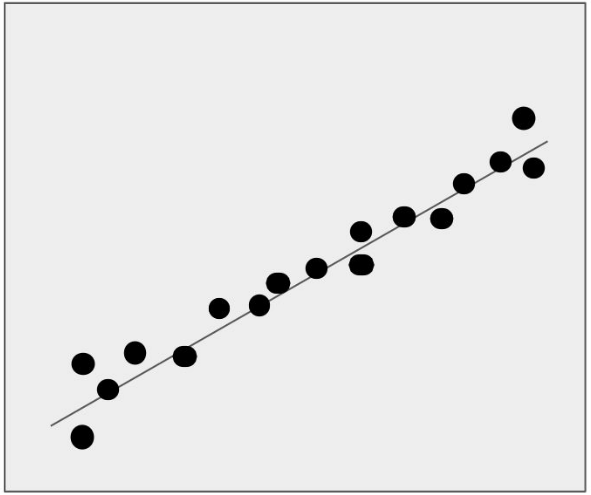


Case 2

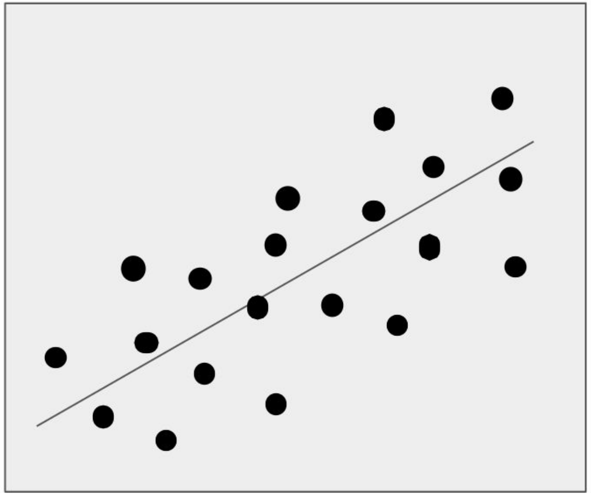


Standard Deviation?

All you need to care about is …. VARIANCE

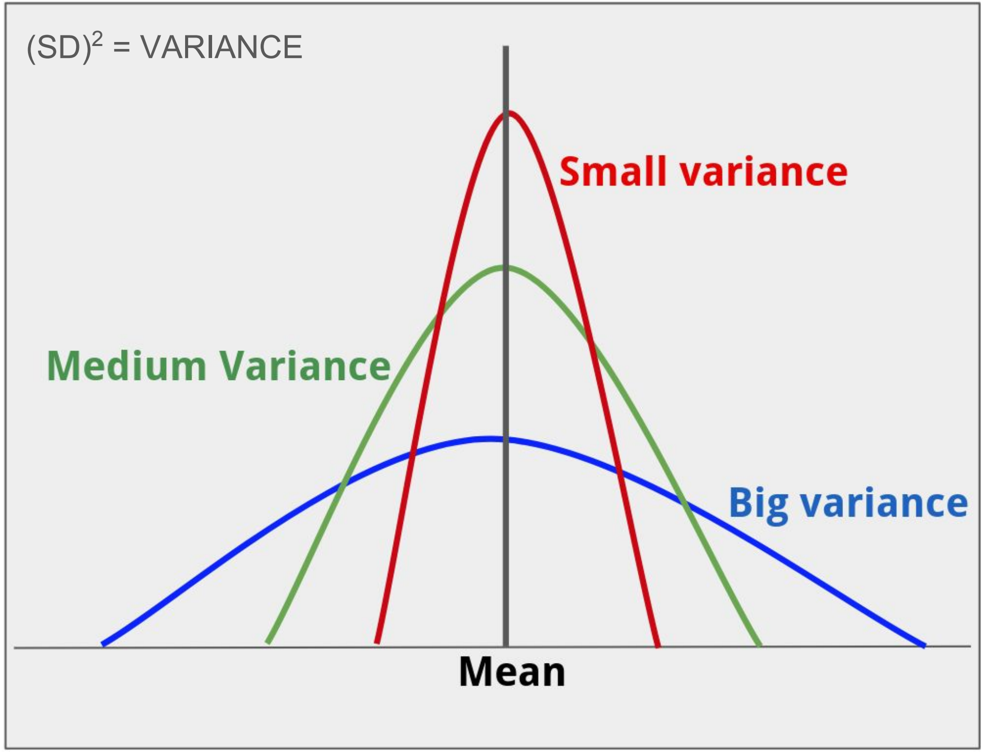


Small Variance



Big Variance

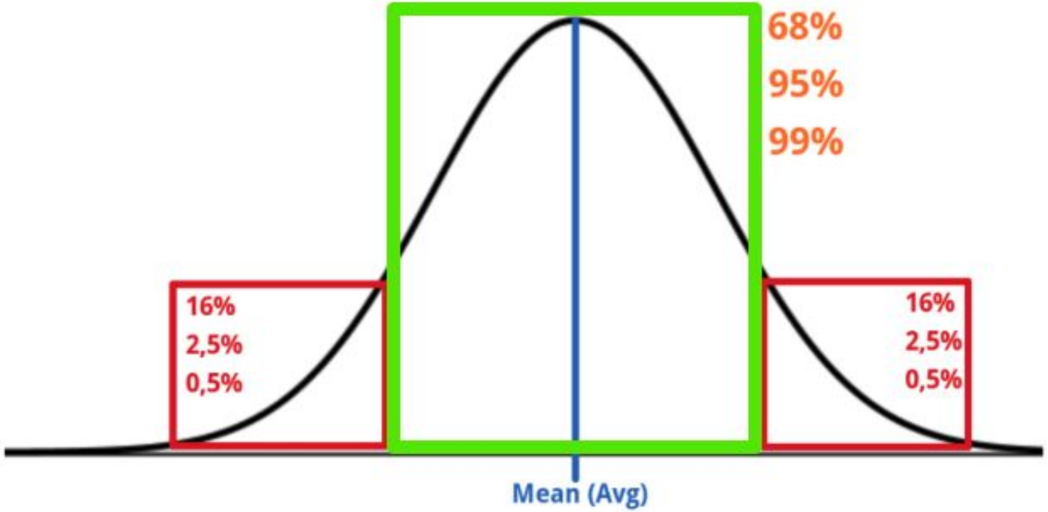
Graphically

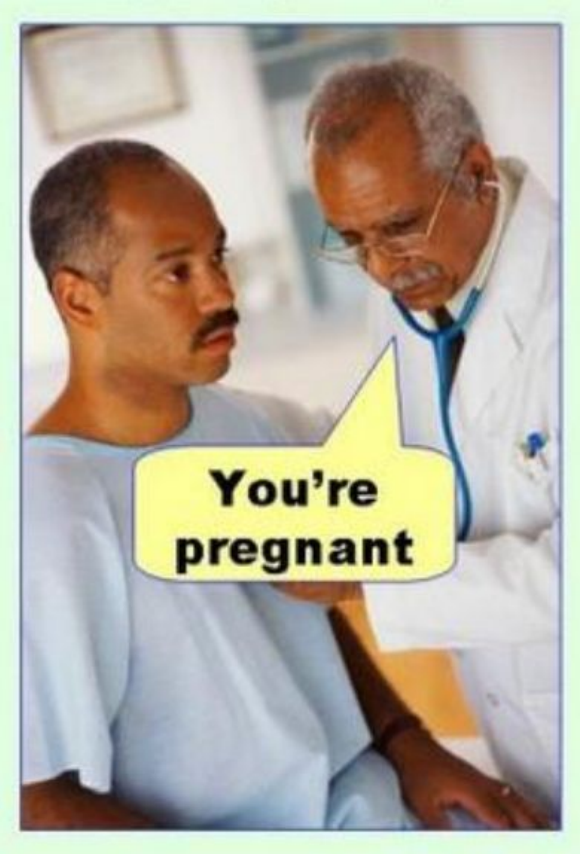


Stat. Significance
(or how to avoid false positive results)

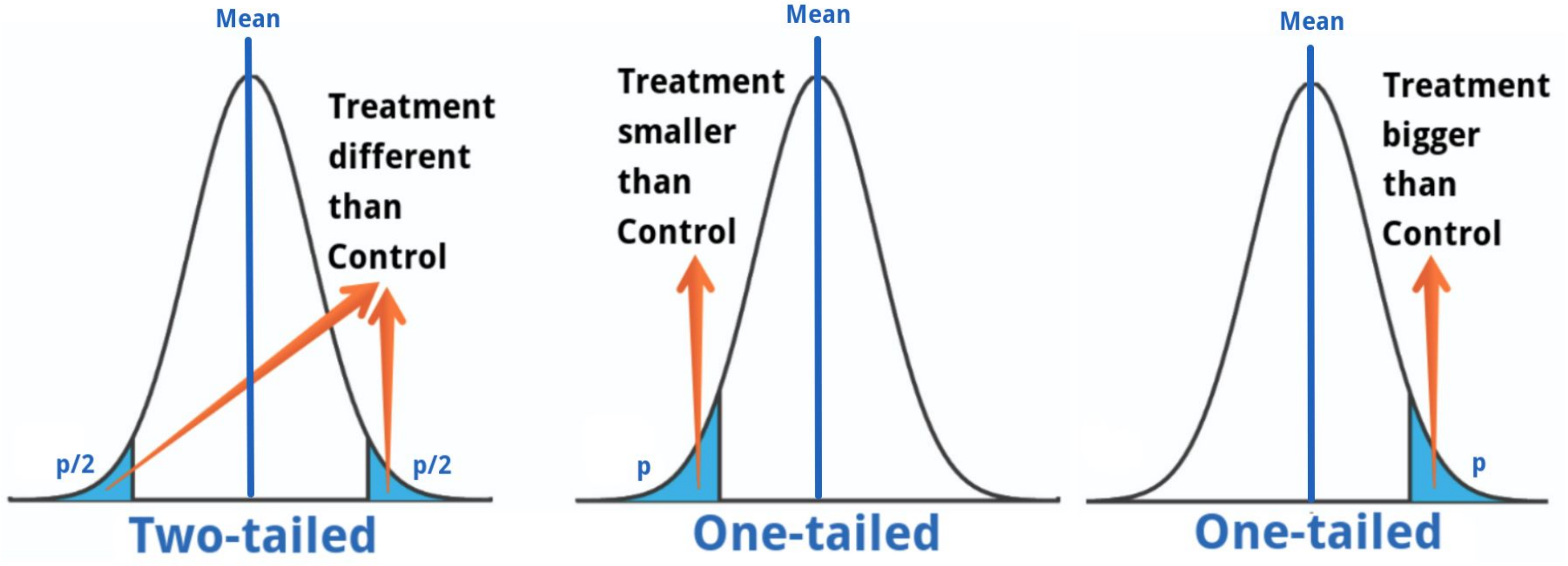
P-value is probability to have false positive results (P = 0.05 is a 5% prob. of having false positive results, due by randomness)

Acceptable Significance level = 0.95 (95%), 0.99 (99%)





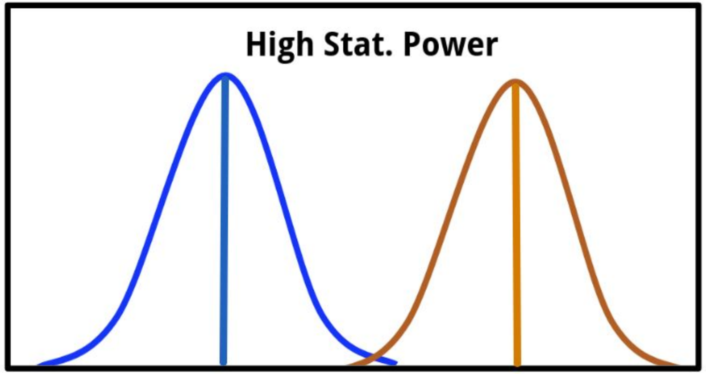
One and Two-tailed Test and P-value position

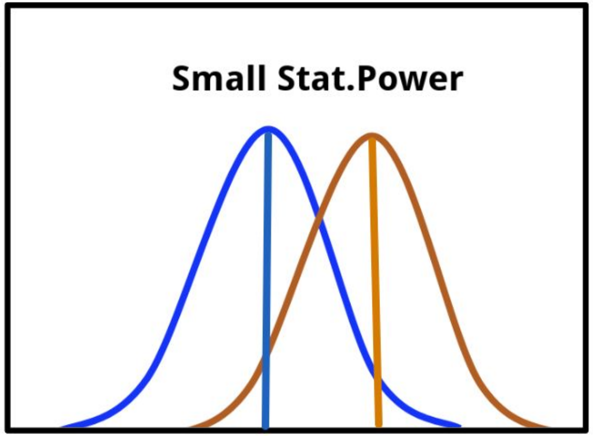

Mean
Treatment different than Control
p/2
p/2
Two-tailed
Mean
Treatment smaller than Control
p
One-tailed
Mean
Treatment bigger than Control
p
One-tailed

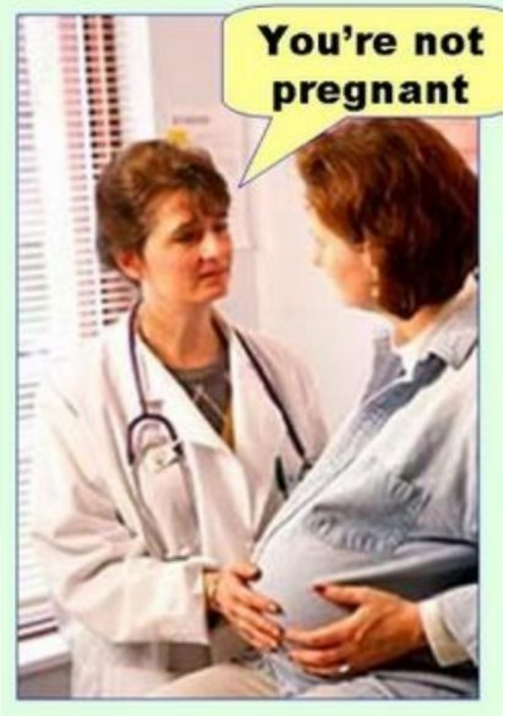
Statistical Power

Percent of the time the minimum effect size will be detected, assuming it exists.

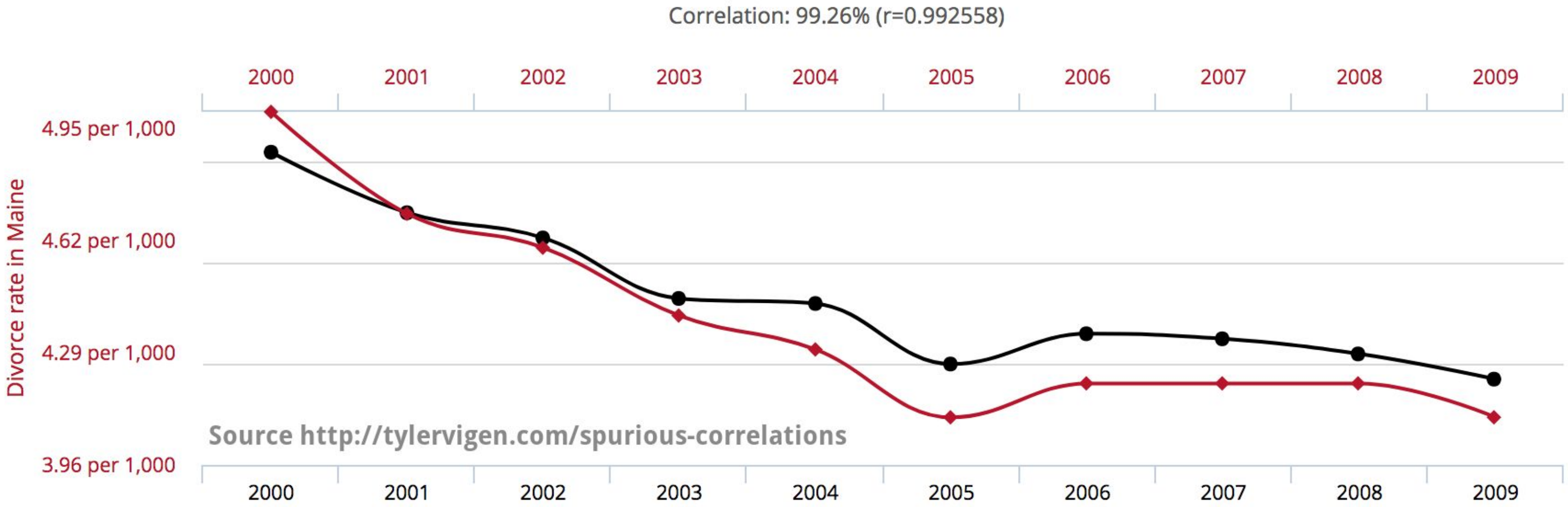
Higher Stat. Power means less probability to have false negative results, an option to increase the Star. Power is increase your sample size (more people).

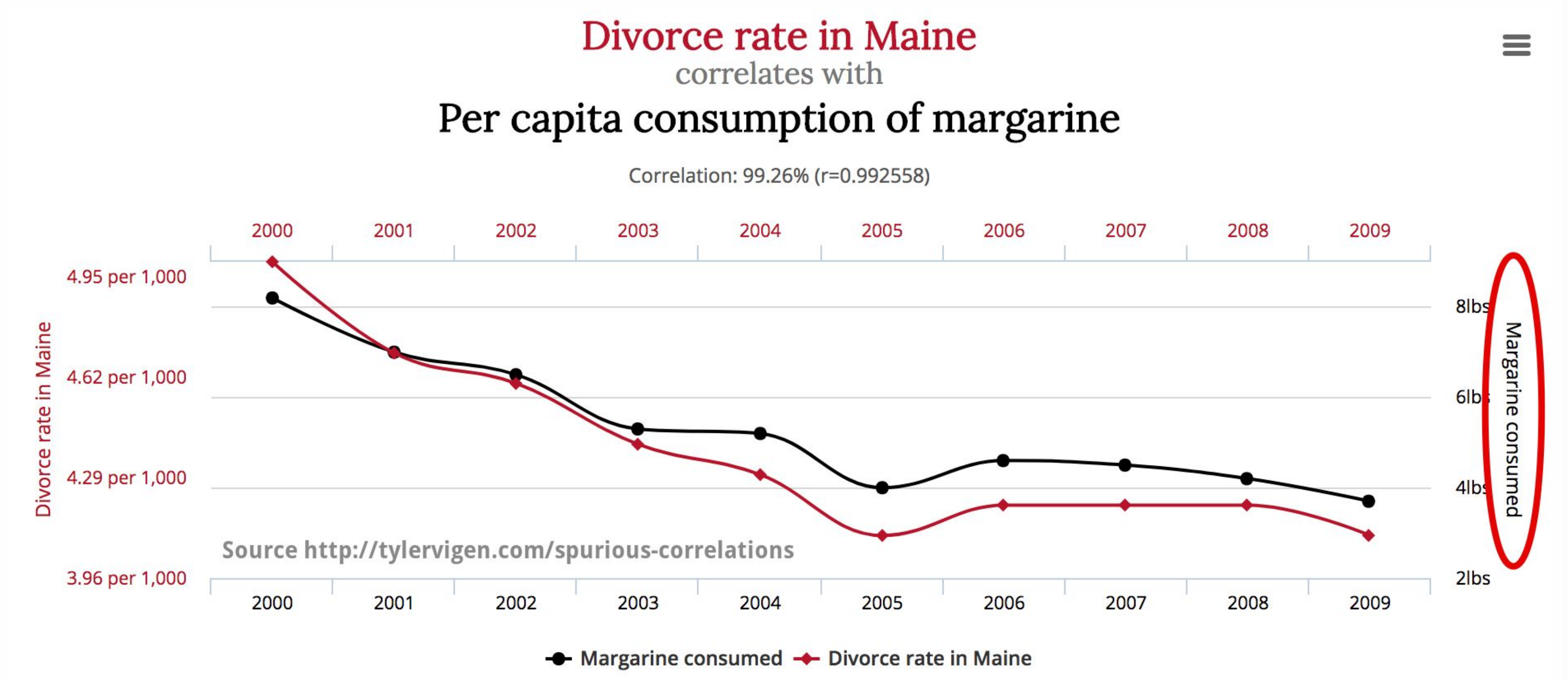






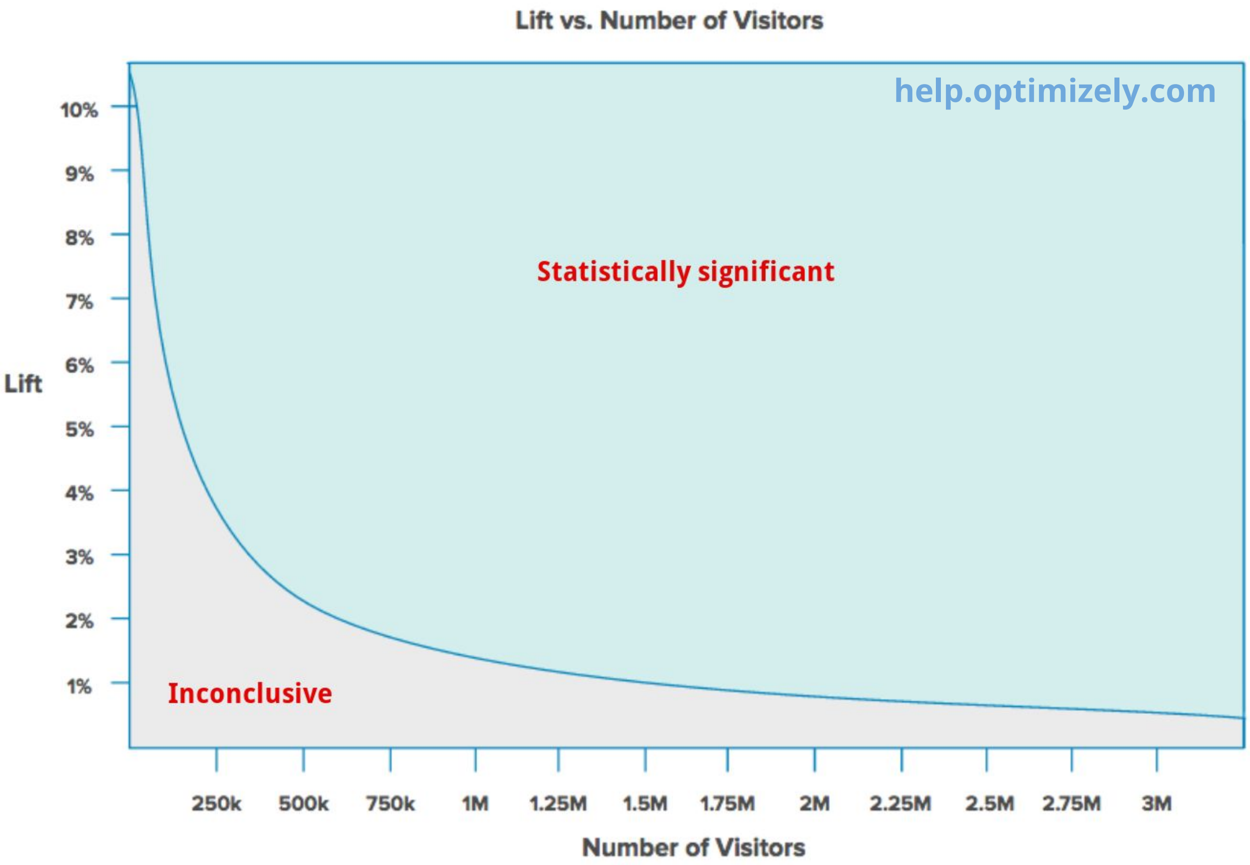
Correlation and importance of the Context





BUT ... not always is so easy to recognize!!

Minimum Detectable Effect



Question: How many people do you think you need to measure a 0.5% difference between your control and your treatment group?

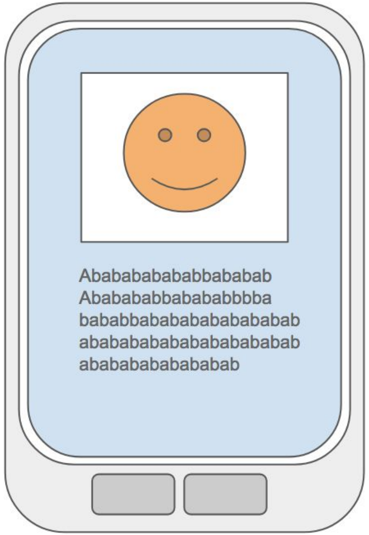
Sample size and Running time

~~Natural groups~~ (natural tendencies could affect the measure)

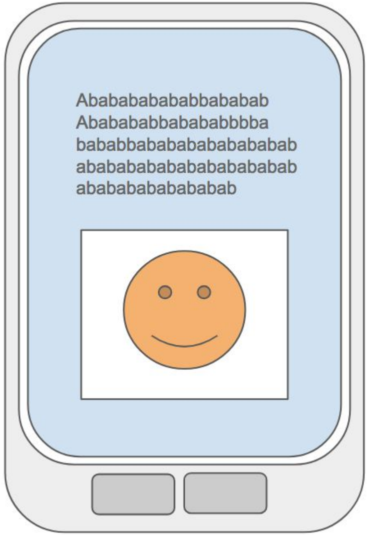
Random groups

You have	**You need**	
LESS days	MORE users	(and vice versa)
SMALL min.eff	MORE users	(and vice versa)
SMALL min.eff	MORE days	(and vice versa)

Total vs Averages (Means)



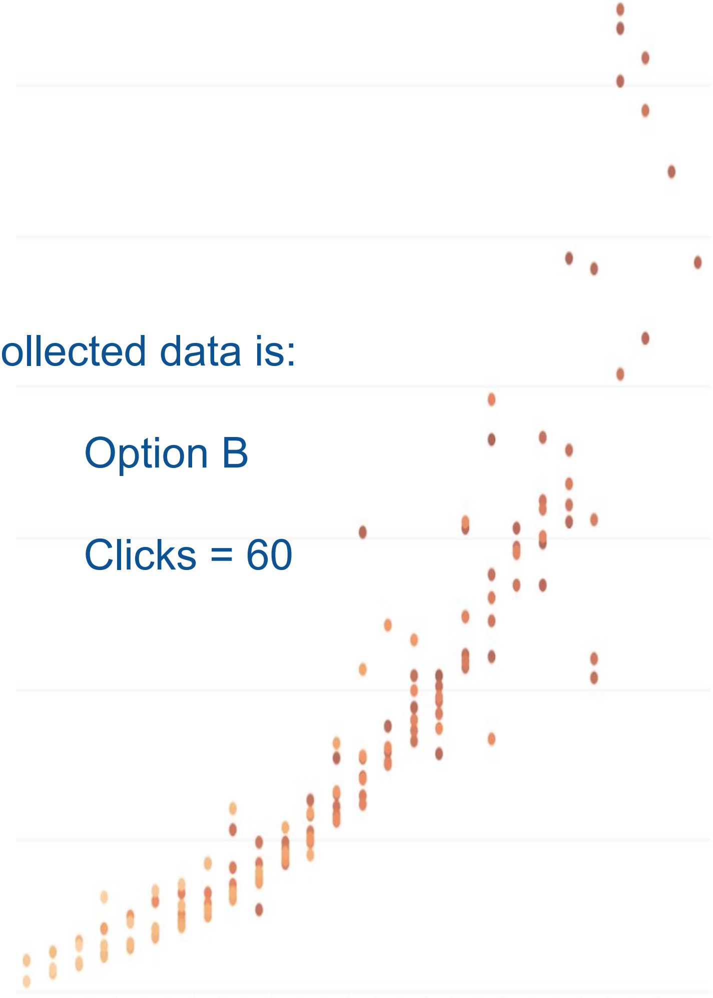
Option A



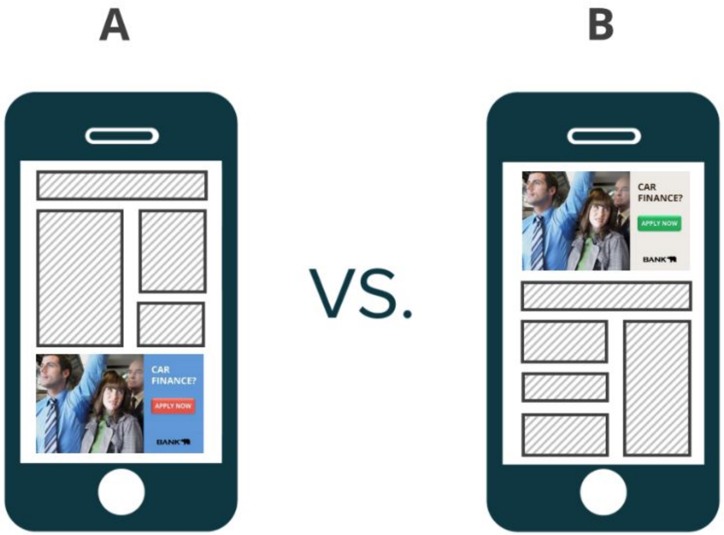
Option B

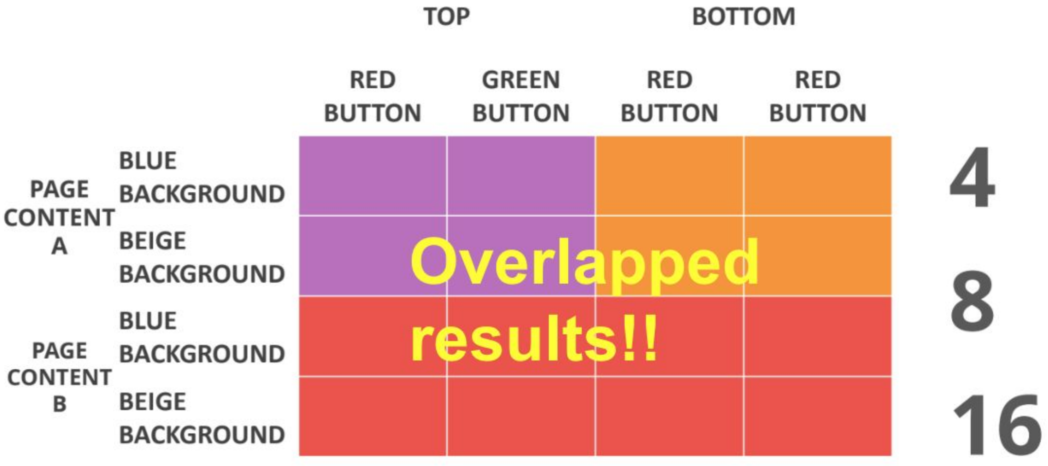
In 3 days your collected data is:

Option A	Option B
Clicks = 50	Clicks = 60



Keep it Simple





TOP

BOTTOM

RED BUTTON

GREEN BUTTON

RED BUTTON

RED BUTTON

PAGE CONTENT A

BLUE BACKGROUND

BEIGE BACKGROUND

PAGE CONTENT B

BLUE BACKGROUND

BEIGE BACKGROUND

Overlapped results!!

4

8

16

Source "Experiment Design, for behavioural Interventions" Facebook talk, April 26th 2018.

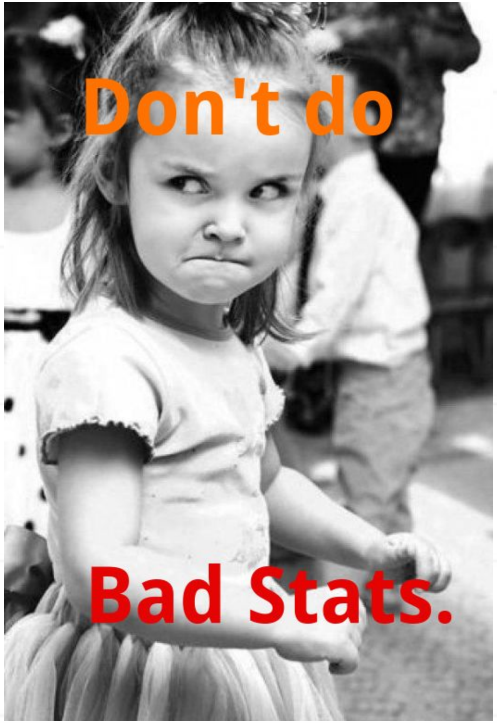
Do your research! (avoid spend resources)

Your “*Don’t do it*” List

1- Hypothesis: **Have one** (even exploratory tests need it)

2- Avoid overlapping tests, your experiment must be the **ONLY** difference between your Control and Treatment groups.

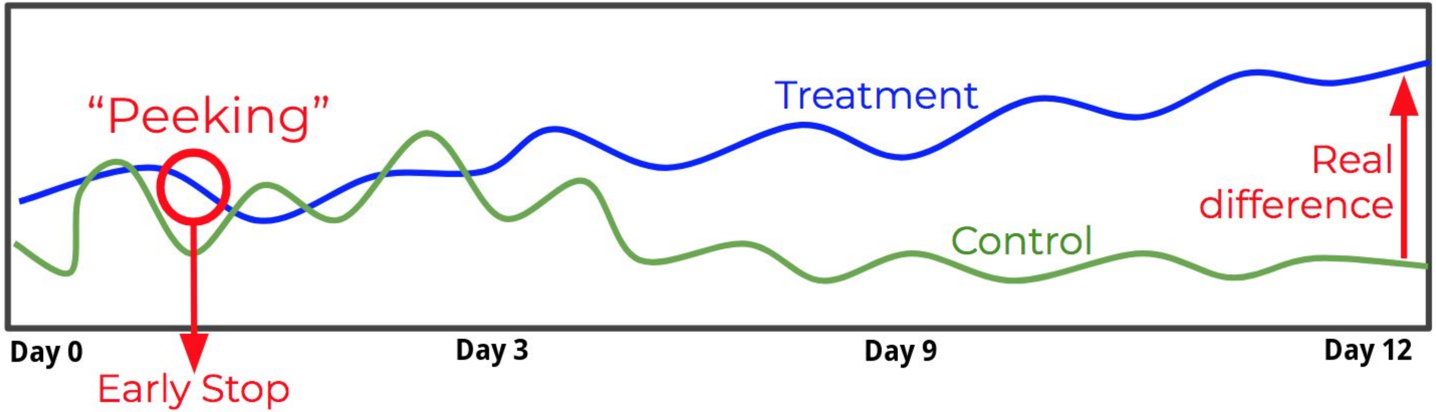
3- Stop you test early = **Ruining** you test



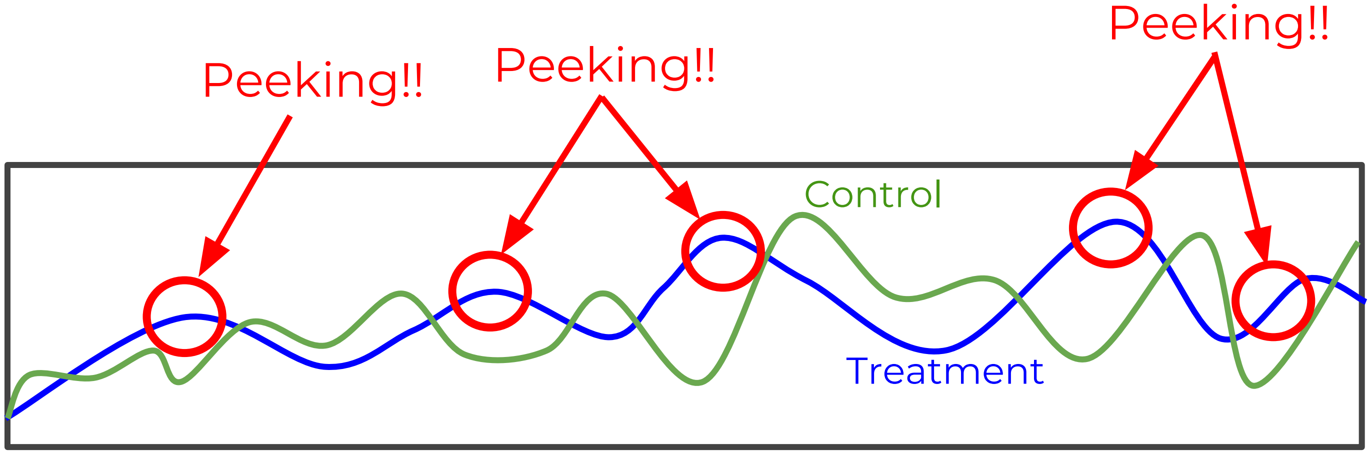
What happens when you stop your test early?

Stat. Significance is reached at some point, because is not a constant value especially at the beginning, so could be produced by random.

Never forget your % of false results!! (depends on your p-value)



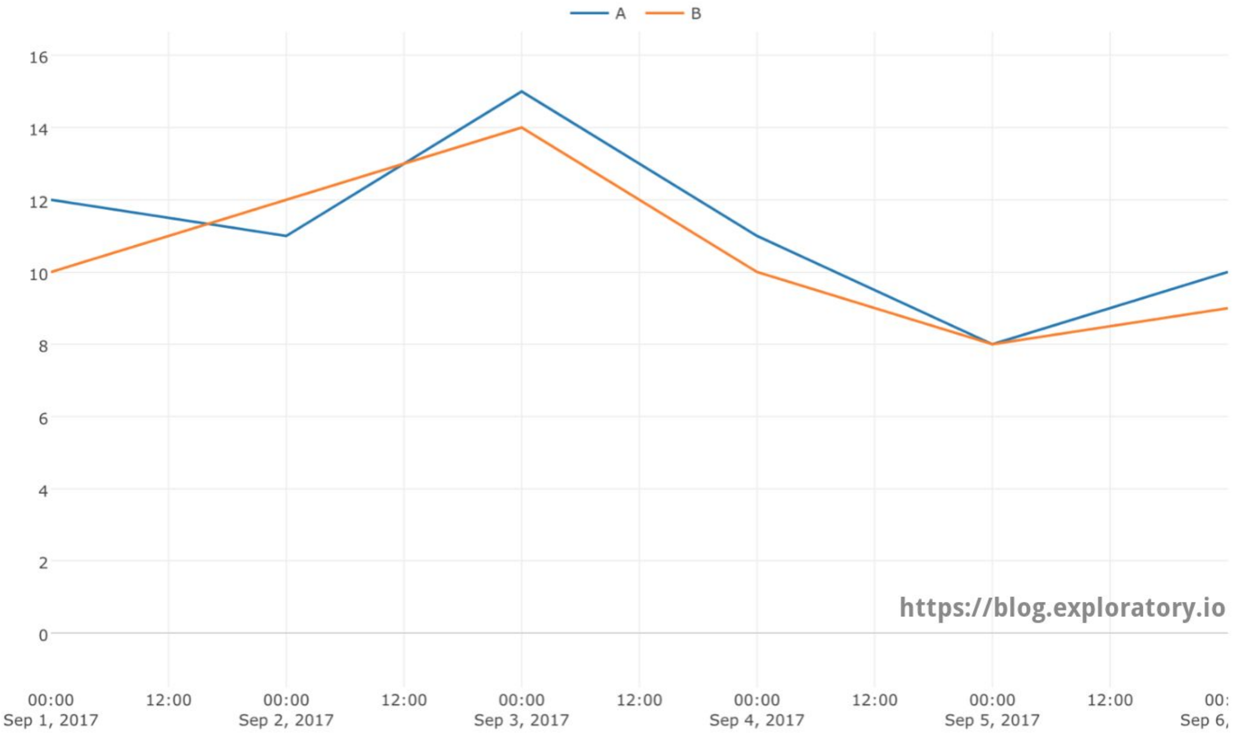
The Peeking Problem



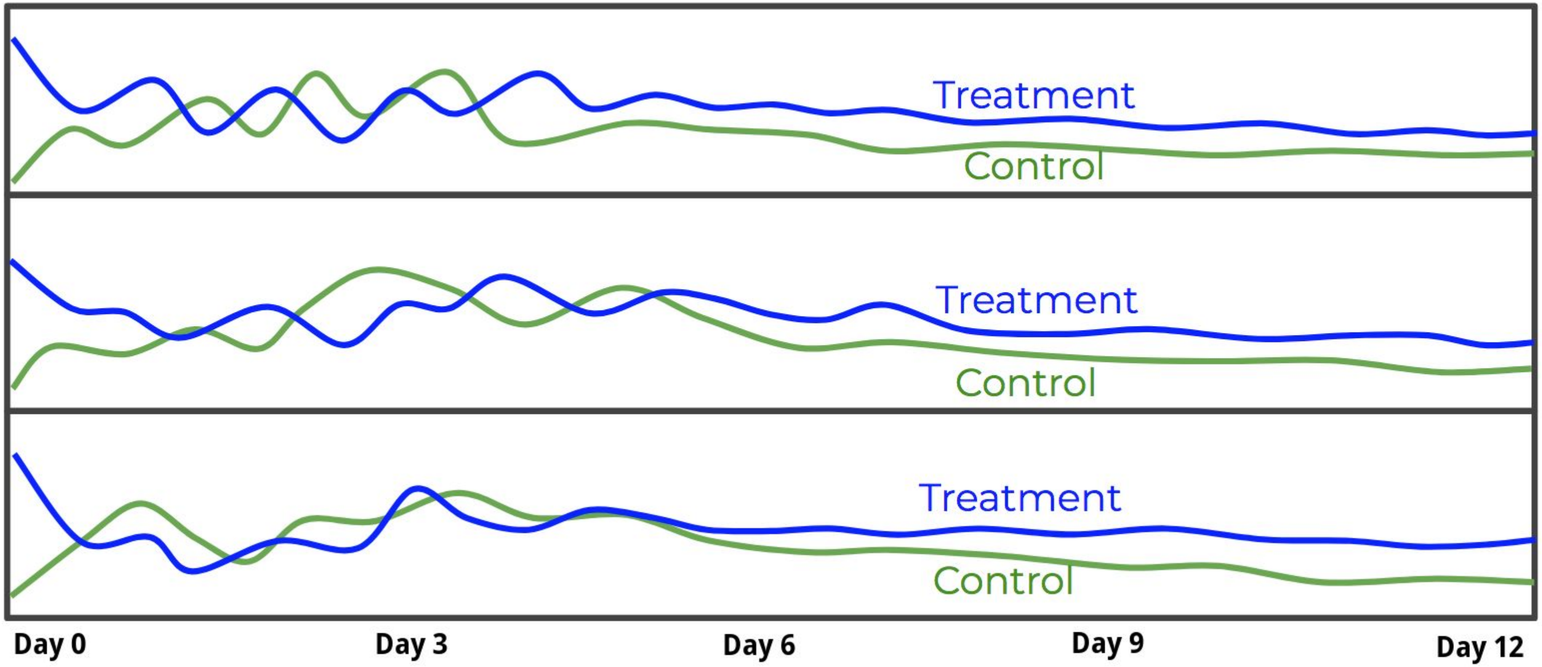
If your test don’t show results could be because:

1- Small sample (small effects will have bigger impacts in small samples)
2- There is not a statistical difference.

Reality: Almost all A/B test won’t produce huge gains



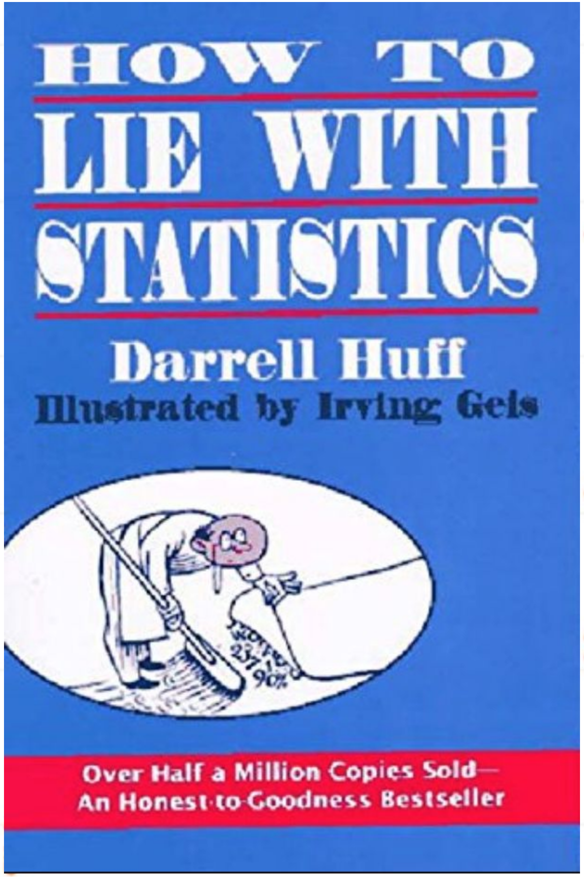
Some useful tests: ROC Curve Analysis, Bayesian A/B Test, Group Sequential Analysis, etc.

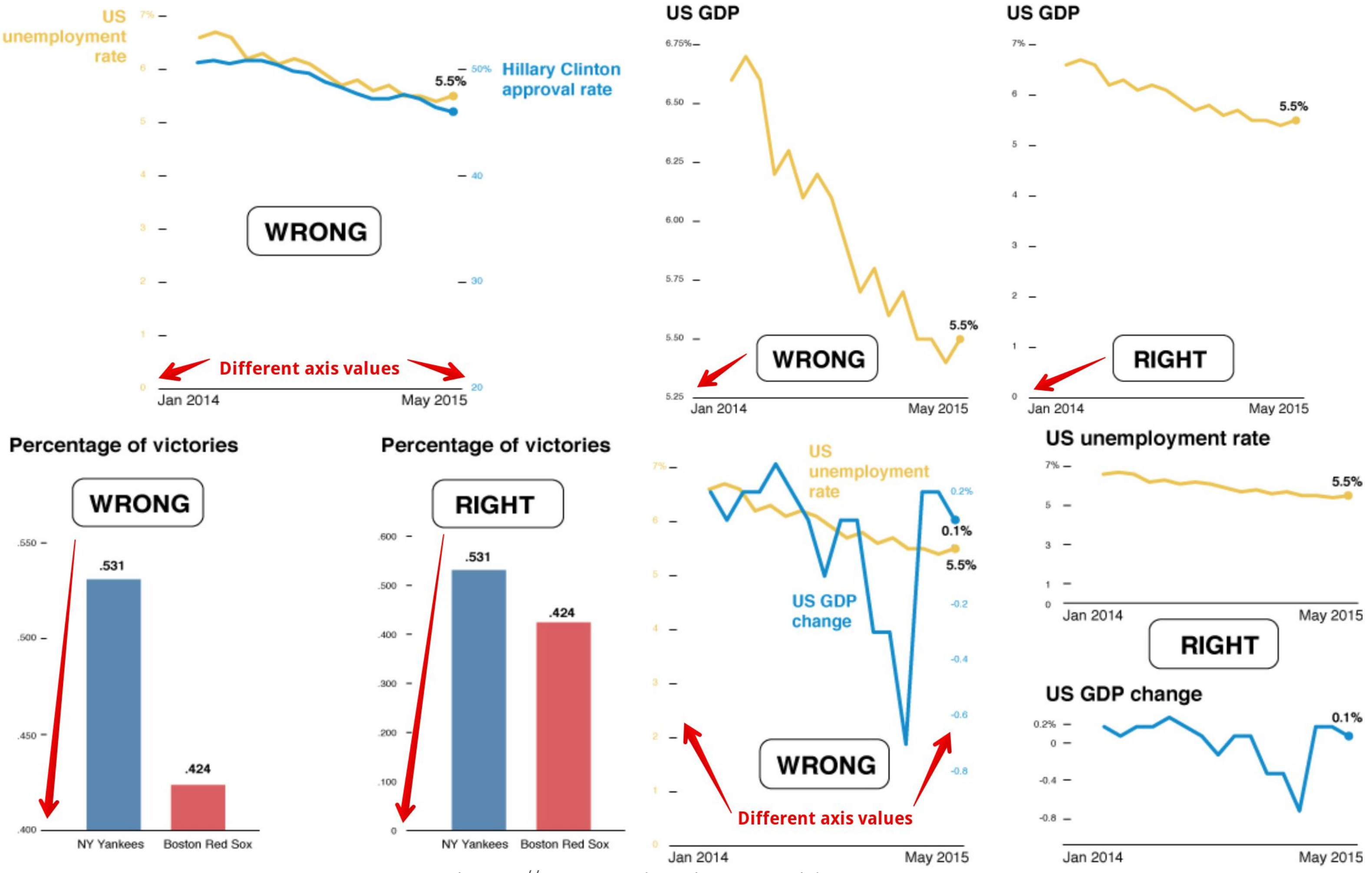


A good A/B test is really a collection of tests.
More important is to be sure the improvement will be permanent.

BONUS:
How to lie with Statistics

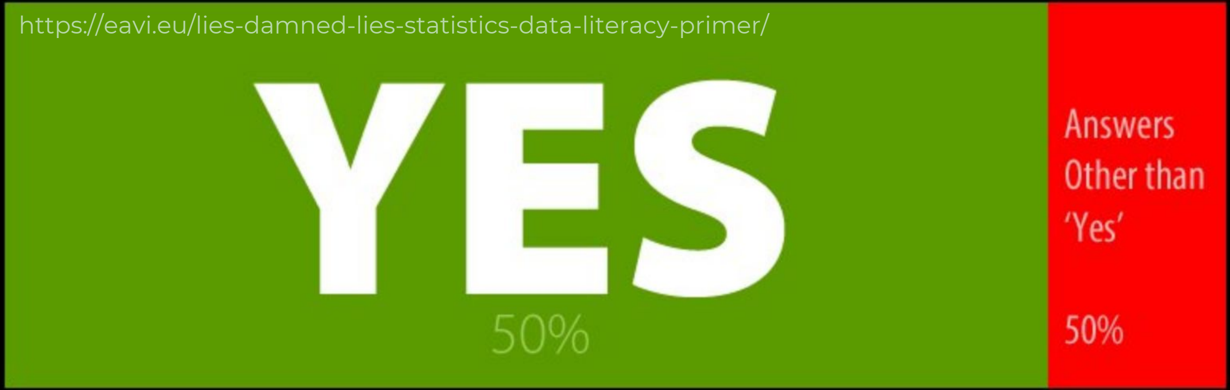
How to Lie with Statistics is a book written by Darrell Huff in 1954 presenting an introduction to statistics for the general reader. Not a statistician, Huff was a journalist who wrote many "how to" articles as a freelancer (Wikipedia)


HOW TO LIE WITH STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

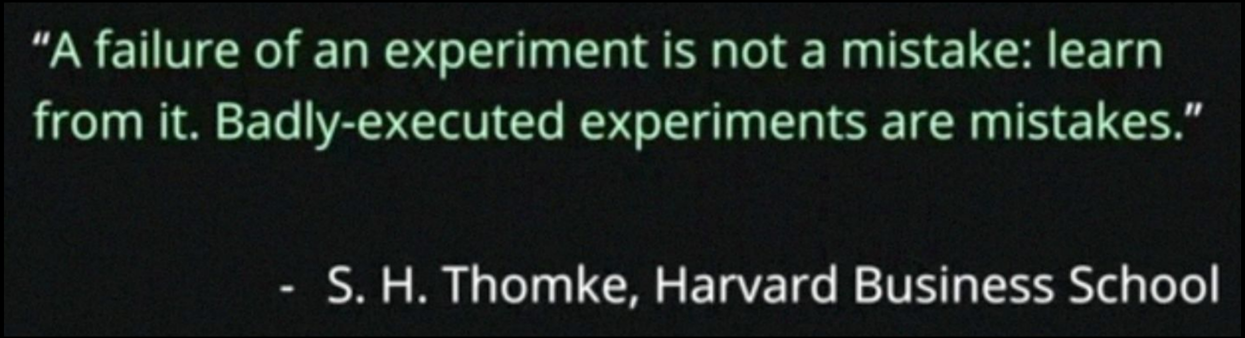


Source https://news.nationalgeographic.com

Should you avoid using misleading data visualisation just to support your argument?



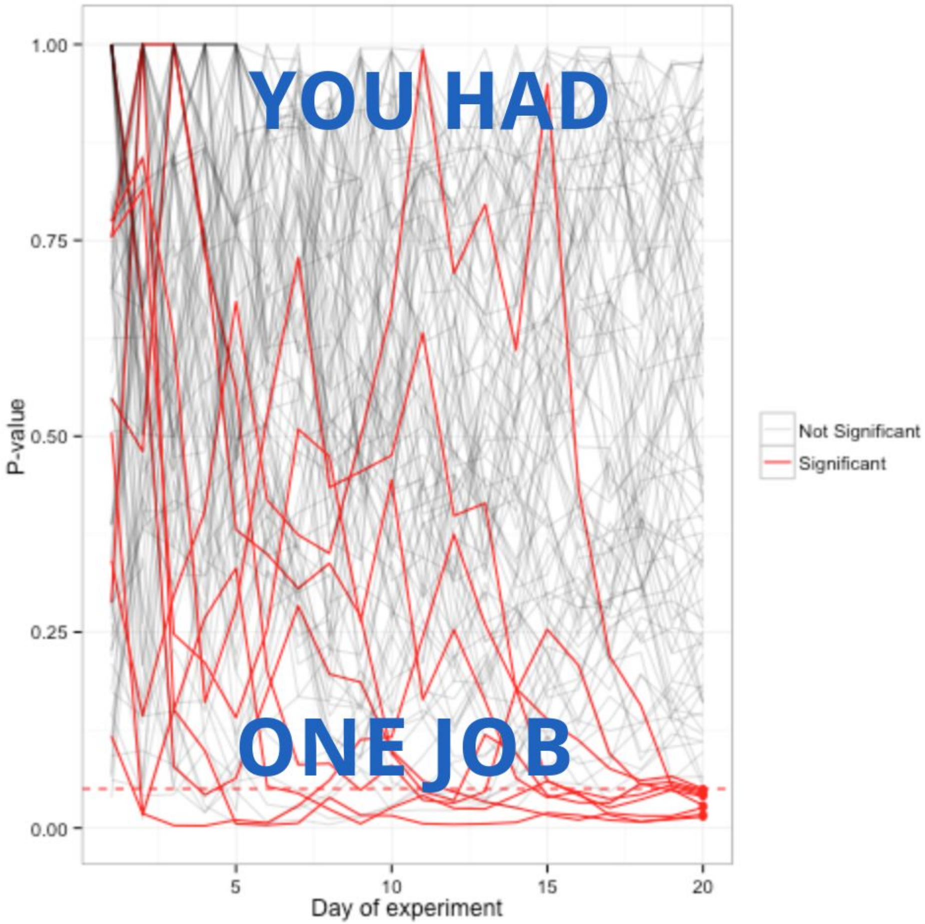
https://eavi.eu/lies-damned-lies-statistics-data-literacy-primer/



"A failure of an experiment is not a mistake: learn from it. Badly-executed experiments are mistakes."

- S. H. Thomke, Harvard Business School

Thanks!


YOU HAD
ONE JOB
1.00
0.75
0.50
0.25
0.00
P-value
5
10
15
20
Day of experiment
Not Significant
Significant

END