Some things that you should know about

A/B Testing

A dozen short lessons

Sept 2016

Goals for today

- Discuss basic ideas of A/B testing: no equations*
- Share some lessons that may be new to even experienced A/B testers

**almost*

Contents

Introduction

Pirates

Suppose a pirate challenges you to bet on a coin toss

A safe bet?

Heads = the pirate wins
Tails = you win

You’re a savvy gambler - what if the coin is weighted?!

A safe bet?

To placate you, the pirate tosses the coin 100 times to prove it is fair

A safe bet?

To placate you, the pirate tosses the coin 100 times to prove it is fair

The toss is heads 55 / 100 times. Should you worry?

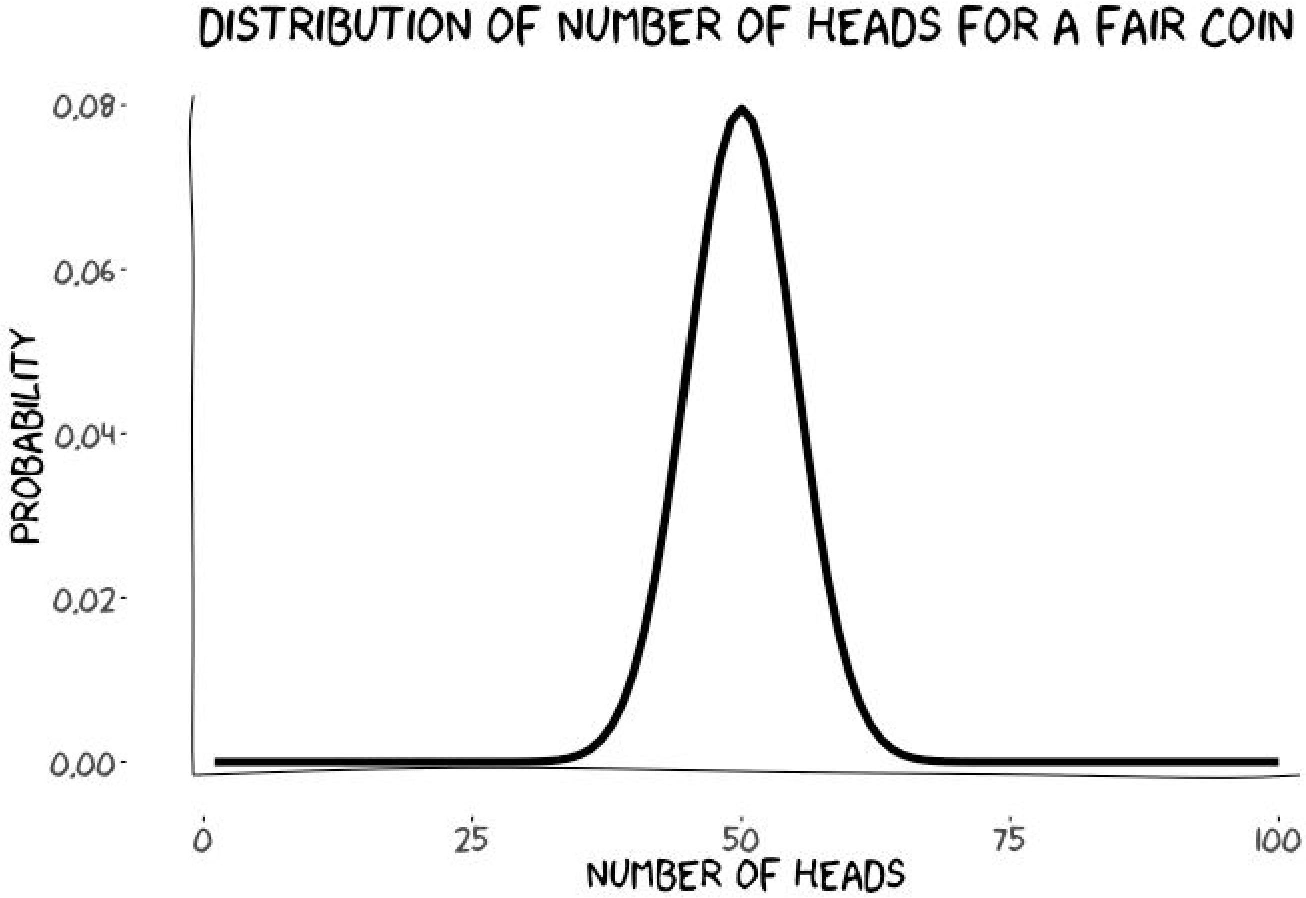
A safe bet?

To placate you, the pirate tosses the coin 100 times to prove it is fair

The toss is heads 55 / 100 times. Should you worry?

What if there were 95 / 100 heads?

A safe bet?


DISTRIBUTION OF NUMBER OF HEADS FOR A FAIR COIN
PROBABILITY
0.08
0.06
0.04
0.02
0.00
0
25
50
75
100
NUMBER OF HEADS

Coin tosses

Should you worry?

- 55/100: Not necessarily.
  - But you might want him to toss the coin some more!
- 95/100: Yes, run matey!

Introduction:
P-values and Confidence Intervals

P-values

Null hypotheses

A null hypothesis is our skeptical, default belief.

Null hypotheses

A null hypothesis is our skeptical, default belief.

Example: the pirate coin is fair

- Is there evidence to the contrary?
- Hypothesis testing is about quantifying that evidence and using it to make decisions

p-values

A p-value is a *measure of the evidence* against the null hypothesis

- How implausible is the data?
- Could it have happened by chance?

p-values

A p-value is a *measure of the evidence* against the null hypothesis

- How implausible is the data?
- Could it have happened by chance?

Two important properties

- P-values range from 0 to 1
- Small p-value -> strong evidence
- Big p-value -> weak or no evidence

Hypothesis testing rule:

Reject the null hypothesis as implausible if p-value < alpha

Hypothesis testing rule:

Reject the null hypothesis as implausible if p-value < alpha

- Often, alpha = 0.05 (just convention)

Hypothesis testing rule:

Reject the null hypothesis as implausible if p-value < alpha

- Often, alpha = 0.05 (just convention)
- If you make decisions with this rule it controls the probability of a “Type I error”
    - *if the null hypothesis is true, there is only a 5% chance you will mistakenly reject it*

Lesson 1:

P-values are not probabilities

P-values are not probabilities

Unfortunately, the p-value is ***not*** the same as the probability that the null hypothesis is true!

P-values are not probabilities

Unfortunately, the p-value is ***not*** the same as the probability that the null hypothesis is true!

- Best interpretation: a measure of evidence

P-values are not probabilities

Unfortunately, the p-value is ***not*** the same as the probability that the null hypothesis is true!

- Best interpretation: a measure of evidence
- This is subtle. In short, you need to make other assumptions for it to even make sense to calculate a probability
- **What is always true**: If the null hypothesis is true, there is only a 5% chance you will reject it by mistake
- We will come back to this

Confidence intervals

Confidence intervals

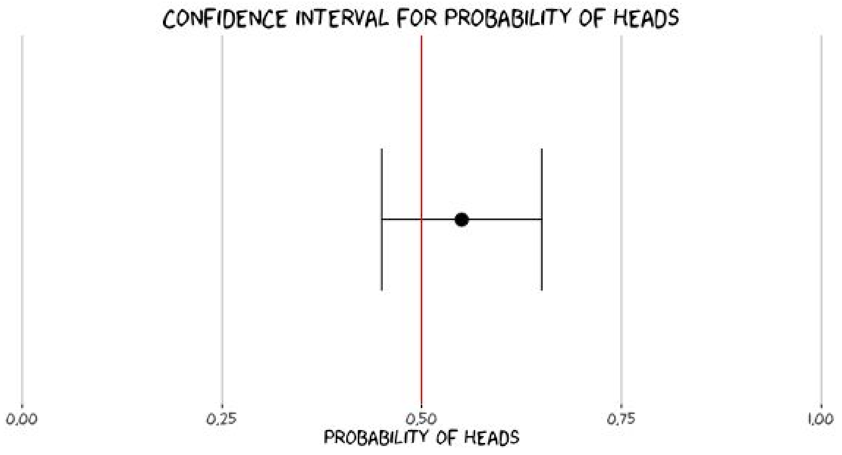
- Confidence intervals summarize the uncertainty in a number by representing it as an interval
- They have a close relationship to p-values and hypothesis testing

Back to our pirate

Confidence intervals for coin tosses

Example:

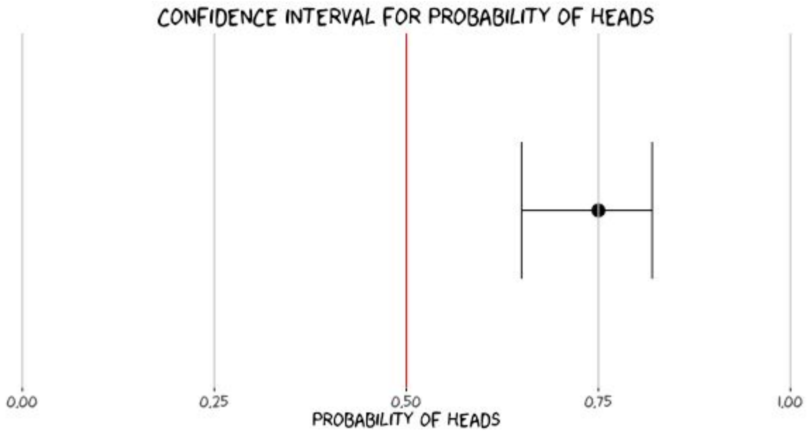
- 55/100 heads
- medium p-value (0.36)
- Confidence interval
  - **[0.45, 0.65]**
- Contains 0.5!



Confidence intervals for coin tosses

Example:

- 75/100 heads
- tiny p-value (almost zero)
- Confidence interval
  - **[0.65, 0.82]**
- Contains 0.5!



Interpretation

- You can assume the **true value** is contained in the confidence interval

- In the long run, you will be right 95% of the time

Interpretation

- You can assume the **true value** is contained in the confidence interval
- In the long run, you will be right 95% of the time
- Checking if 0.5 is contained in the interval is the same as testing the null hypothesis that the coin is fair

Interpretation

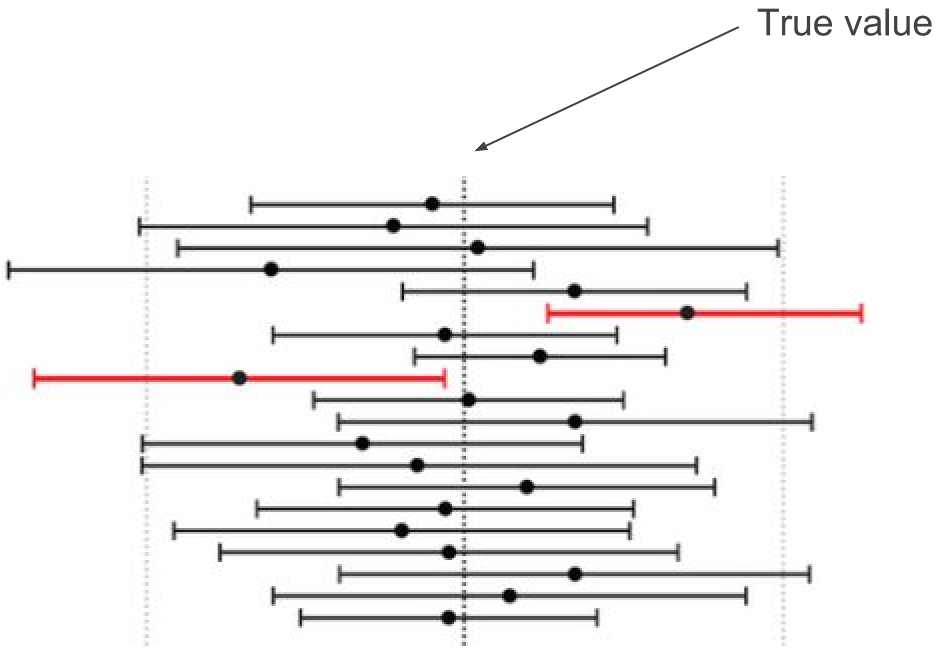
- You can assume the **true value** is contained in the confidence interval
- In the long run, you will be right 95% of the time
- Checking if 0.5 is contained in the interval is the same as testing the null hypothesis that the coin is fair
- Like p-values, this is *not* a probability statement about each confidence interval after the fact - they either contain the true value or they don't!

Interpreting confidence intervals

Imagine repeating the experiment many times

- The confidence interval would “catch” the true value 95% of the time

- But for each experiment, it either contains the true value or it doesn’t



A/B testing

A/B testing is about comparing two groups

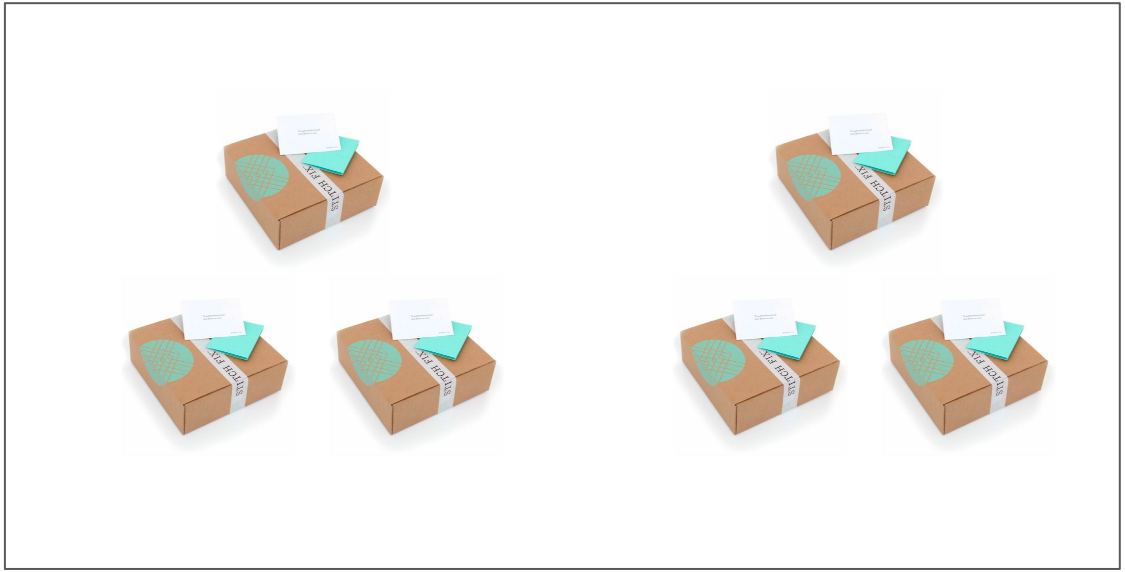
- For example, comparing the impact two different versions of an algorithm

- A/B tests are the gold standard because the randomization eliminates all other sources of difference between the two groups
  - Time
  - Clients
  - Inventory
  - Etc

There are called confounding variables.

A/B testing with fixes

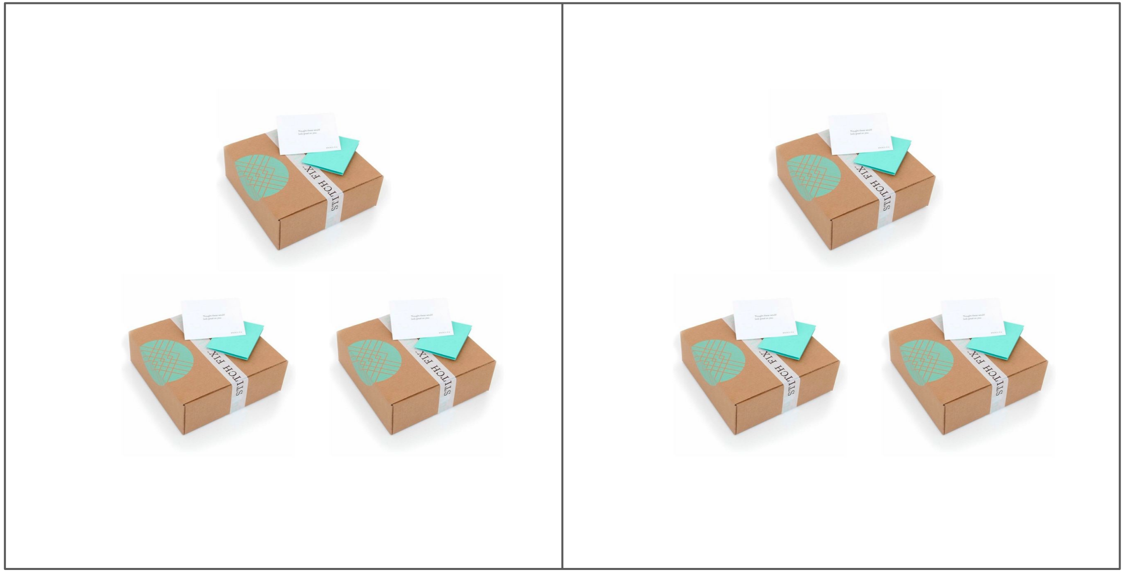
Take a set of fixes to include in the experiment



A/B testing

Randomly divide the shipments into two groups. Because of the randomization they should be the same, on average.

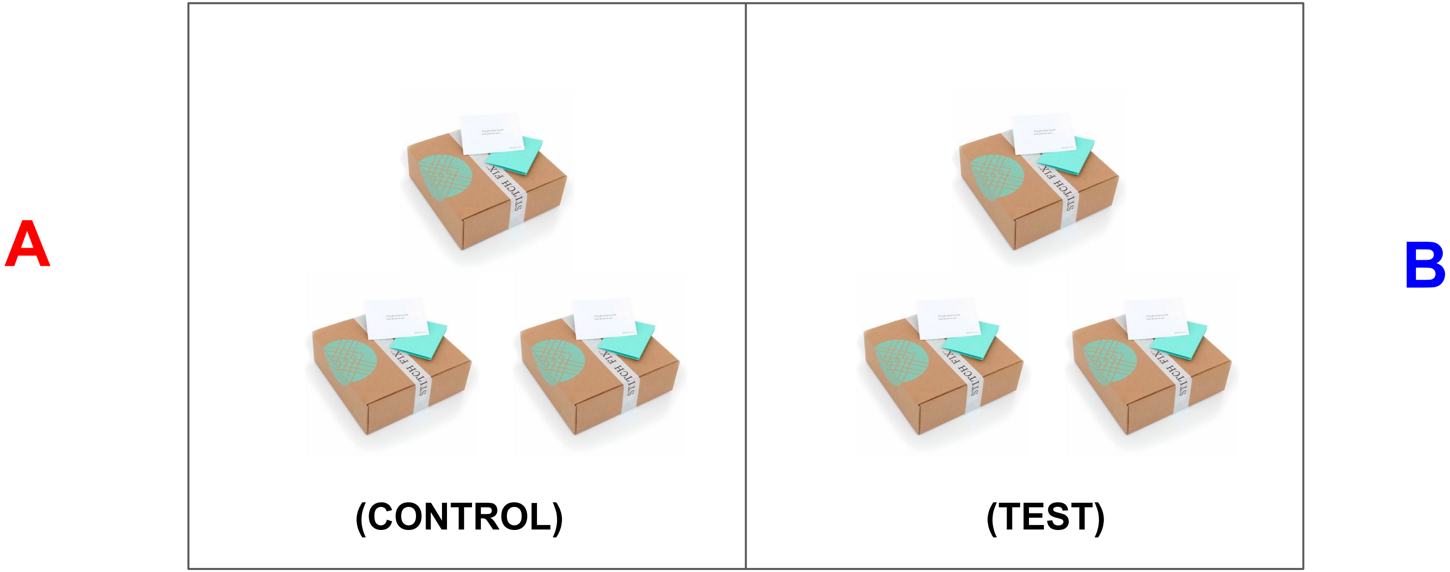
A



B

A/B testing

Now, do something to one of the groups and measure the impact



A real example

Comparing ArgStyle (test) to Argyle (control)

v9-5	v9-0-1	Diff	p Value
112.7338 **(149459)**	111.6794 **(149317)**	1.0544 **(0.4085 to 1.7003)**	0.00138

- Shipments randomly assigned to test or control
- Null hypothesis: no difference between algorithms
- Confidence interval for difference in AOV
    - [$0.40, $1.70]

http://abtestament.vertigo.stitchfix.com/experiments/26/results

Lesson 2:

Randomization is everything

Randomization

Randomization is the foundation of A/B testing

- **Intuition**: we want to balance the cells so that they are a fair comparison

Randomization

Randomization is the foundation of A/B testing

- **Intuition**: we want to balance the cells so that they are a fair comparison

- Suppose we want to compare two styling algorithms
  - Autoship clients tend to have better outcomes
  - To be fair, we should have an equal proportion of autoship clients in each cell

Randomization

- We know about autoship status. What about all the other latent differences we can’t observe?

- Randomization helps ensure balance across these hidden variables

- And, it allows us to mathematically describe the balance

Lesson 3:
“Everything else” is not a control

“Everything else” is not a control

Key assumption from randomization for 50/50 tests:

P(fix X is in *test*) = **P**(fix X is in *control*)

Any fix in the experiment must have been equally likely to have been assigned to test or to control

Examples

Example of a bad control

- Experiment: Apply a new algorithm for first fixes
- Cells
  - Control: fixes 1+
  - Test: first fixes

Example of a bad control

- Experiment: Apply a new algorithm for first fixes
- Cells
  - Control: fixes 1+
  - Test: first fixes
- The problem
  - First fixes may be different!
  - The chance of being in the control cell varies by fix
    - Probability = 1 for fixes 1+
    - Probability = 0 for first fixes

Examples

Example of a bad control

- Experiment: Send a marketing message with facebook
- Cells:
  - Test: 50% of clients who have facebook (randomly selected!)
  - Control: Everybody else

Examples

Example of a bad control

- Experiment: Send a marketing message with facebook
- Cells:
    - Test: 50% of clients who have facebook (randomly selected!)
    - Control: Everybody else
- The problem
    - Clients without facebook are all in the control!

Two-step allocation

Two ensure a proper control, employ a two-step allocation

- First, select *all* fixes that will be in the experiment (either in test *or* in control)

Two-step allocation

Two ensure a proper control, employ a two-step allocation

- First, select *all* fixes that will be in the experiment (either in test *or* in control)

- Then, within this selected group, randomly assign them to test and control

Two-step allocation

Two ensure a proper control, employ a two-step allocation

- First, select *all* fixes that will be in the experiment (either in test *or* in control)
- Then, within this selected group, randomly assign them to test and control

Life lesson: It is hard to recreate the first step after the fact. It is important to select the fixes for test and control at the same time (and to write them down!)

Lesson 4:
Statistical significance is not practical significance

Good news!

We’ve run a test with NewAlgo!

AOV has increased with a p-value of 0.0001!

Good news?

We've run a test with NewAlgo!

AOV has increased with a p-value of 0.0001!

Time to break out the champagne?

Good news :(

Not so fast!

- All the p-value does is give us confidence that AOV really did increase - that it's not just due to chance

Good news :(

Not so fast!

- All the p-value does is give us confidence that AOV really did increase - that it’s not just due to chance
- It does not mean that the finding is “significant” in practice

Good news :(

Not so fast!

- All the p-value does is give us confidence that AOV really did increase - that it’s not just due to chance
- It does not mean that the finding is “significant” in practice
- Given enough data, even a $0.01 increase could be significant

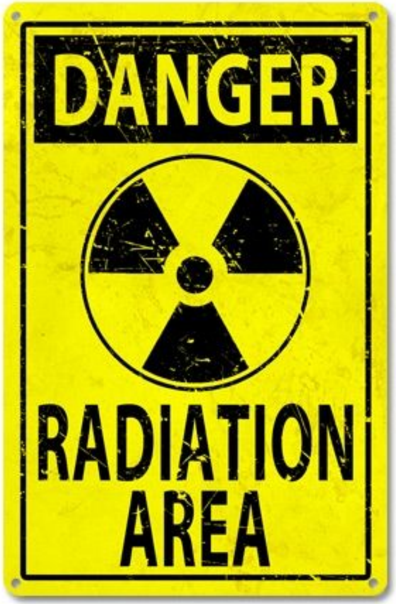
Lesson 5: Statistical fine print matters

A world with many warnings

In life there are some warnings you can probably ignore

A world with many warnings

and some warnings you probably shouldn't



Statistical fine print

(Some of) the statistical “fine print” matters

Statistical fine print

(Some of) the statistical “fine print” matters

- If you “stretch” data when it suits you it will affect the error rate of your decisions

Statistical fine print

(Some of) the statistical “fine print” matters

- If you “stretch” data when it suits you it will affect the error rate of your decisions
- You can choose your own standard of evidence
  - it’s up to you
  - but it will affect how many mistakes you make

Statistical fine print

Some warning signs for lowering the standard of evidence

- *“non-significant improvement”*
- *“trending significant”*
- *“directional”*
- *“almost significant”*

Sometimes this is worse than others - context matters

Some examples

Statistical fine print

Some examples

p = 0.06 - “almost significant”

- You say you test at 0.05, but in practice accept anything < 0.10 as “close enough”
- That’s fine, but it’s a weaker standard of evidence

Statistical fine print

Some examples

p = 0.06 - "almost significant"

- You say you test at 0.05, but in practice accept anything < 0.10 as "close enough"
- That's fine, but it's a weaker standard of evidence

p = 0.80 - "not significant but in the right direction"

- Means almost nothing

Lesson 6:
Early peeking can lead to bad decisions

Early peeking

Suppose you are running an A/B test that will take a month

Early peeking

Suppose you are running an A/B test that will take a month

- During this time p-values will jump up and down as data comes in

Early peeking

Suppose you are running an A/B test that will take a month

- During this time p-values will jump up and down as data comes in
- The temptation to peek can be irresistable!

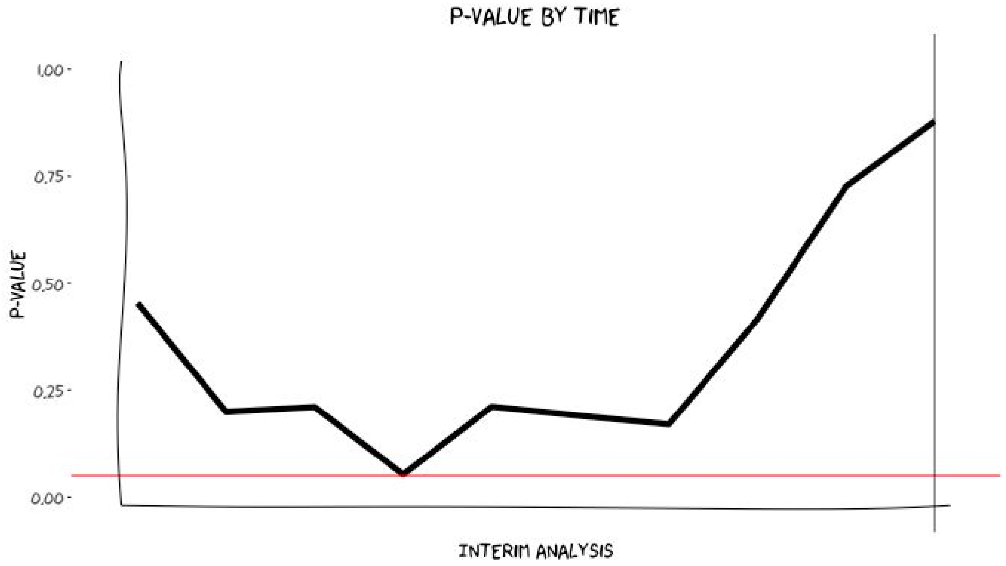
Early peeking

Suppose you are running an A/B test that will take a month

- During this time p-values will jump up and down as data comes in
- The temptation to peek can be irresistable!
- If you check for significance to get “early reads” you are at a much higher risk of a false positive

Up and down

- The probability of the p-value becoming “significant” at some point during the experiment is much higher than 0.05
- **Even if the null hypothesis is true**!



Early peeking ruins your control of type I error

Number of interim Analyses	Type I error if we reject whenever $P < 0.05$
1	0.05
2	0.08
3	0.11
4	0.13
5	0.14
10	0.19
100	0.37

[For more on this topic (source of table)]

Early stopping

Stopping an experiment, or making a decision, early carries additional risk of a making a mistake

There are more advanced techniques that allow for this (ask AA)

Lesson 7:
You have to be careful making multiple comparisons

Type I error rate

A/B testing guarantees are about testing *one* hypothesis at a time

Type I error rate

A/B testing guarantees are about testing *one* hypothesis at a time

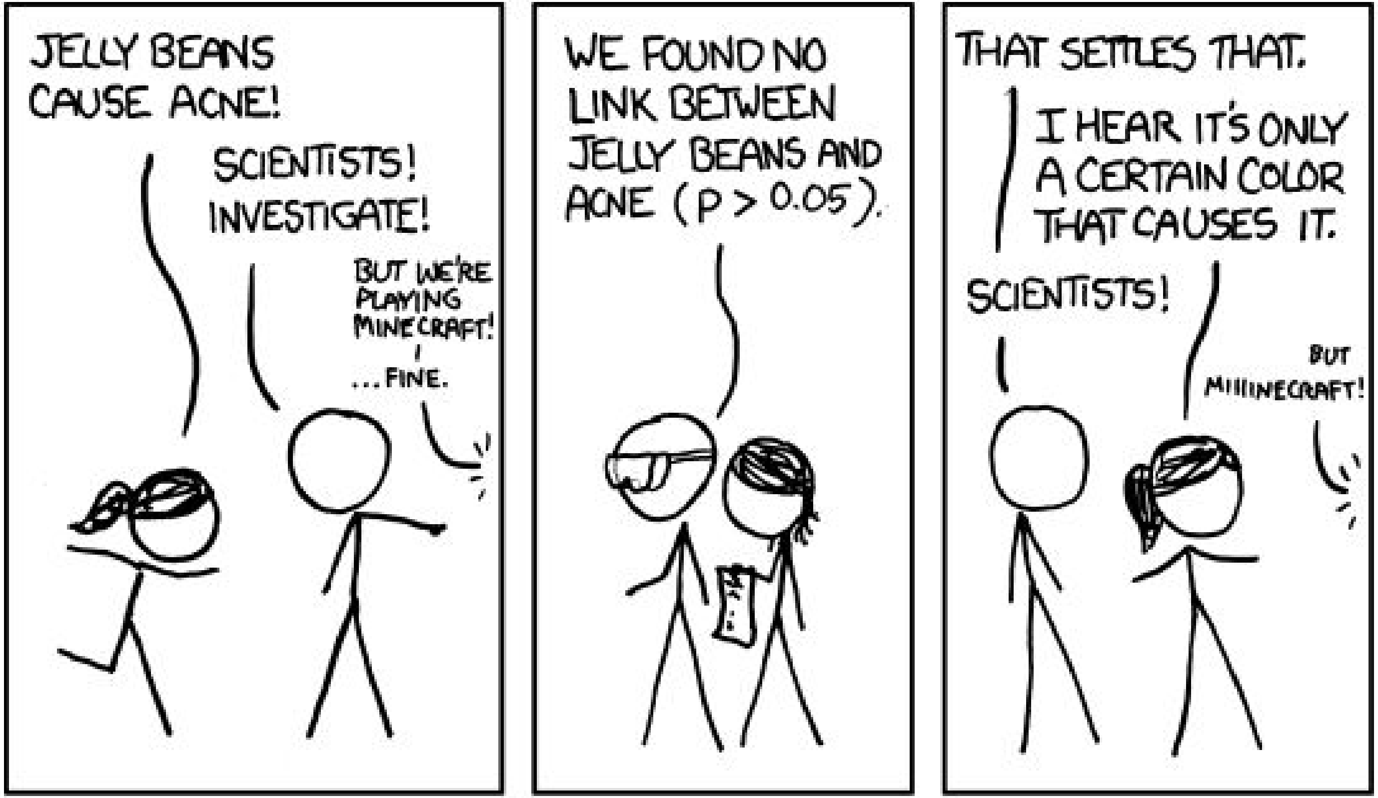
- If you test many hypotheses at a time, the chance of *at least one false positive* can be much higher than 0.05

Type I error rate

A/B testing guarantees are about testing *one* hypothesis at a time

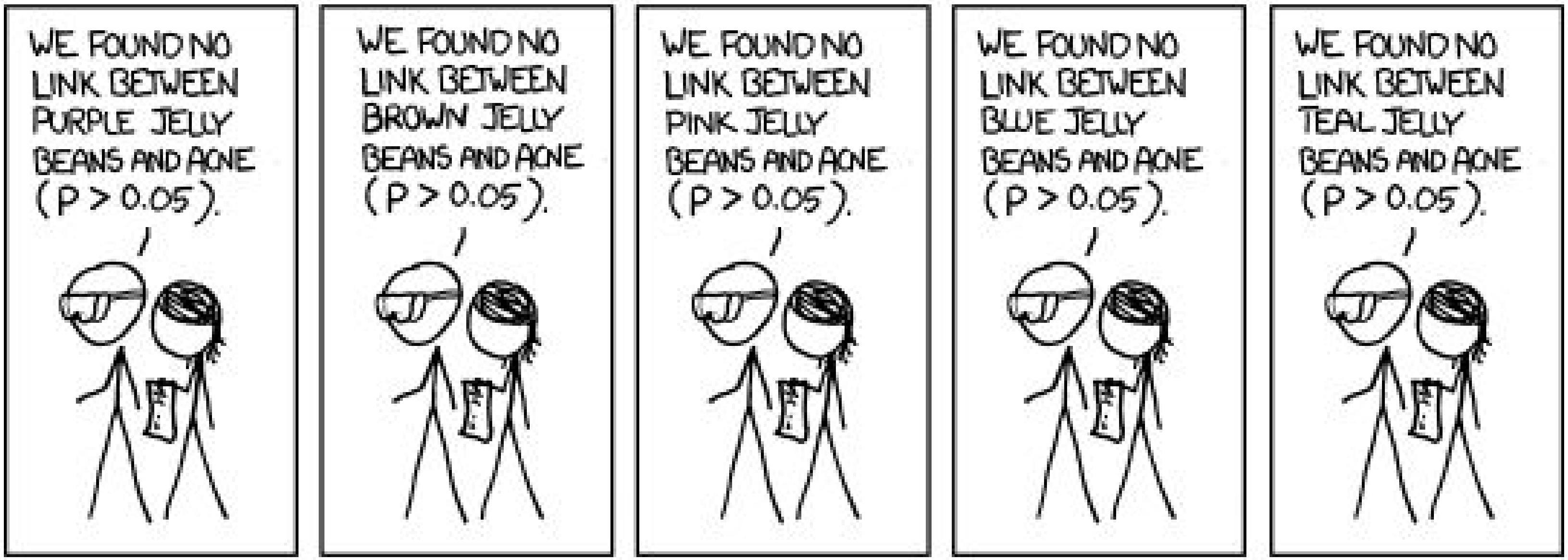
- If you test many hypotheses at a time, the chance of *at least one false positive* can be much higher than 0.05
- **Example**: If you test 20 independent true null hypotheses with alpha = 0.05, you expect one false positive just by chance!

Multiple testing with subsets

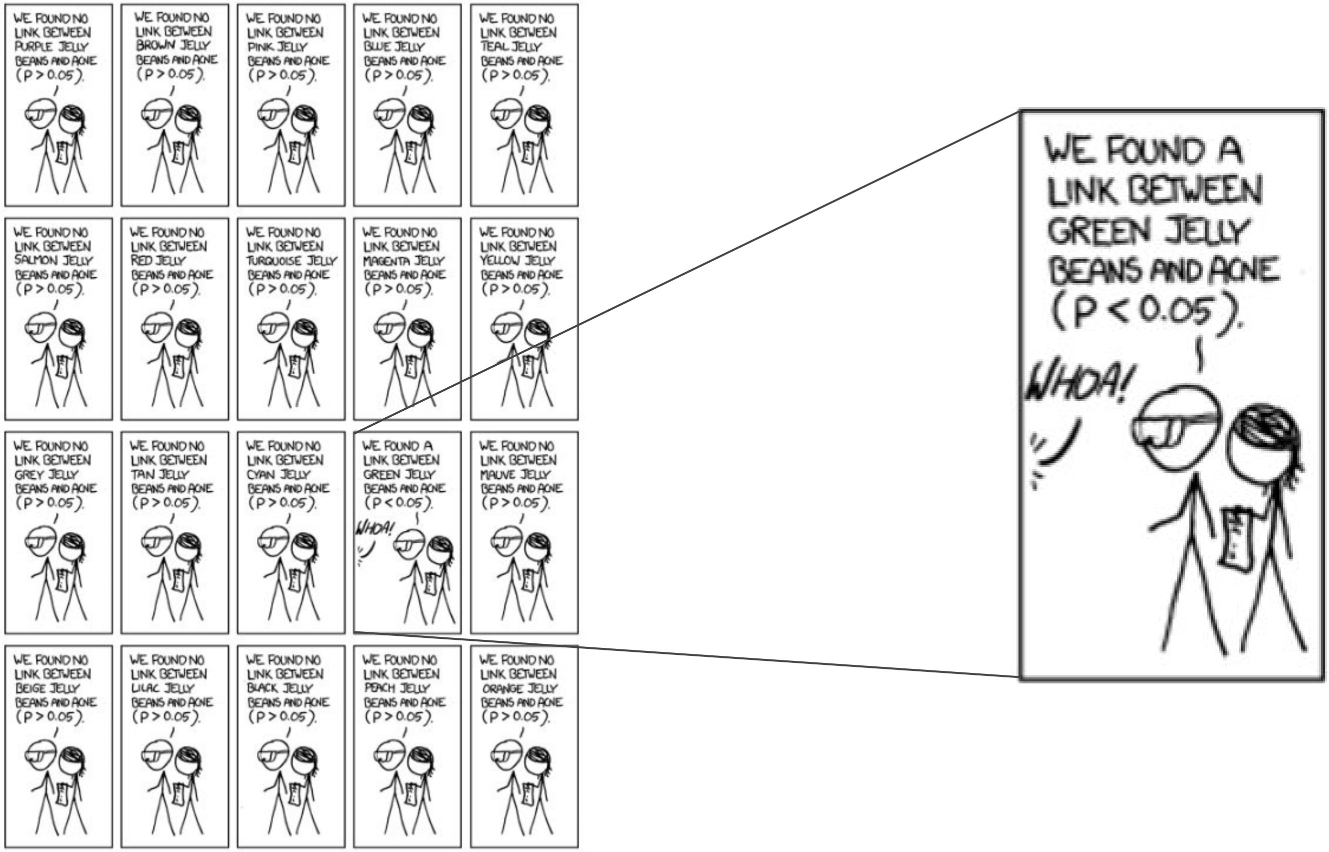


https://xkcd.com/882/

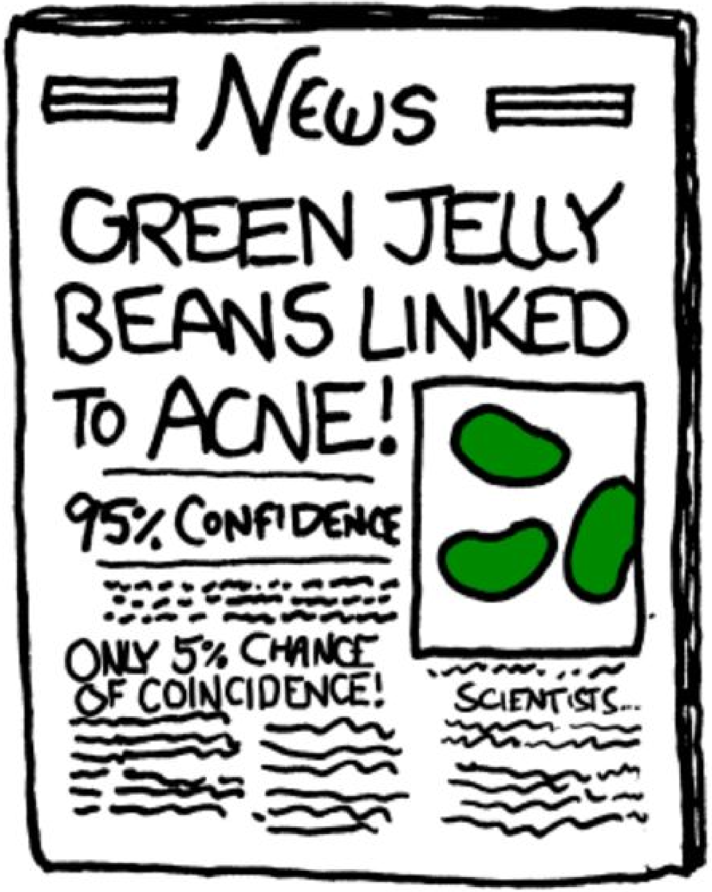
Multiple testing with subsets



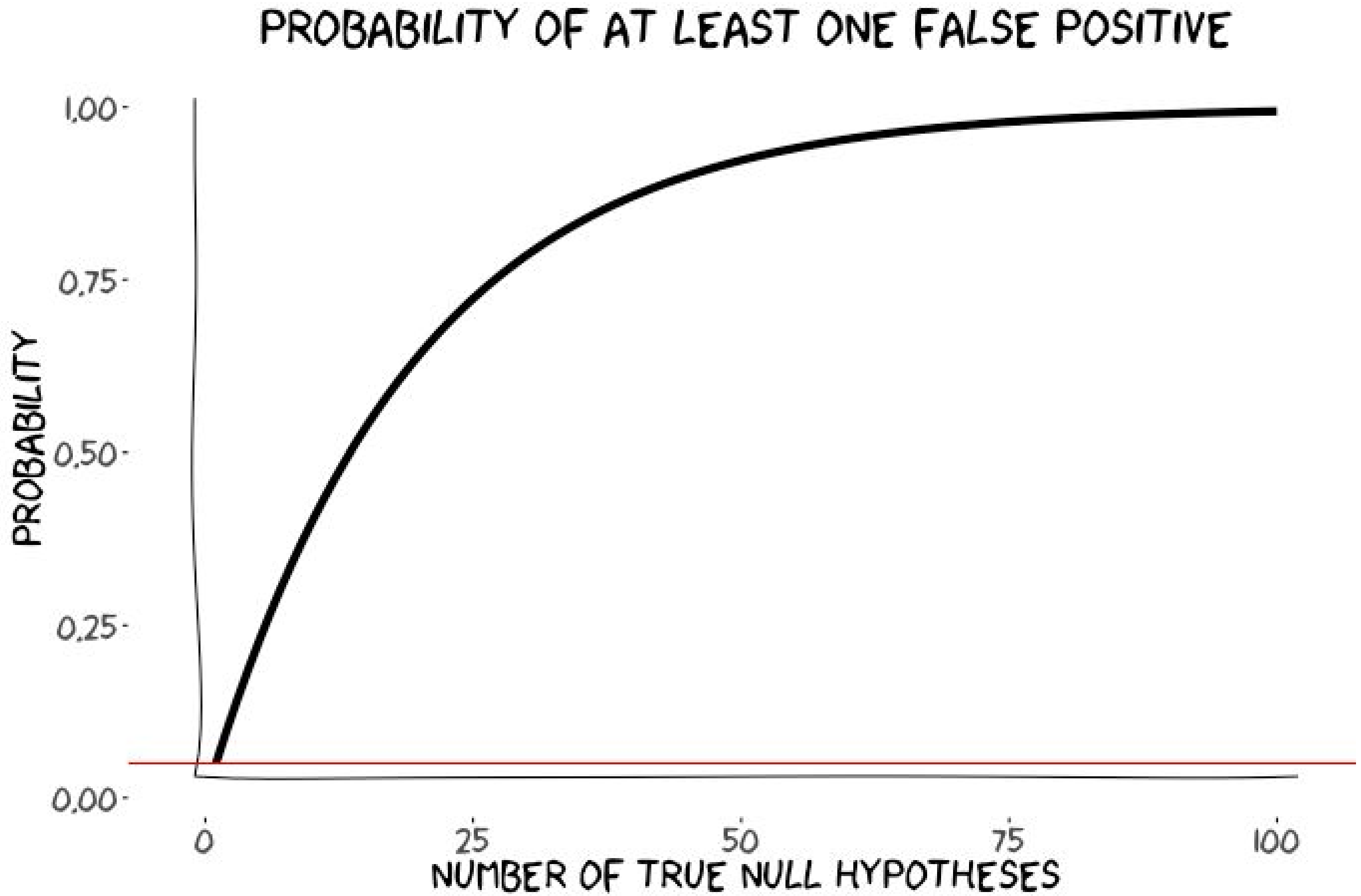
Multiple testing with subsets



Multiple testing with subsets



Assuming independence



Testing multiple hypotheses

What if you want to test multiple hypotheses?

- You have to raise your standard of evidence
- There are statistical methods for doing this
    - See “family-wise error rate” and “false discovery rate”

Lesson 8:
You have to have enough power to find what you’re looking for

Type II error

So far, we've talked about type I errors:

falsely claiming an effect when there isn't one

Type II error

So far, we've talked about type I errors:

falsely claiming an effect when there isn't one

Complementary idea: type II errors

failing to detect an effect when there is one

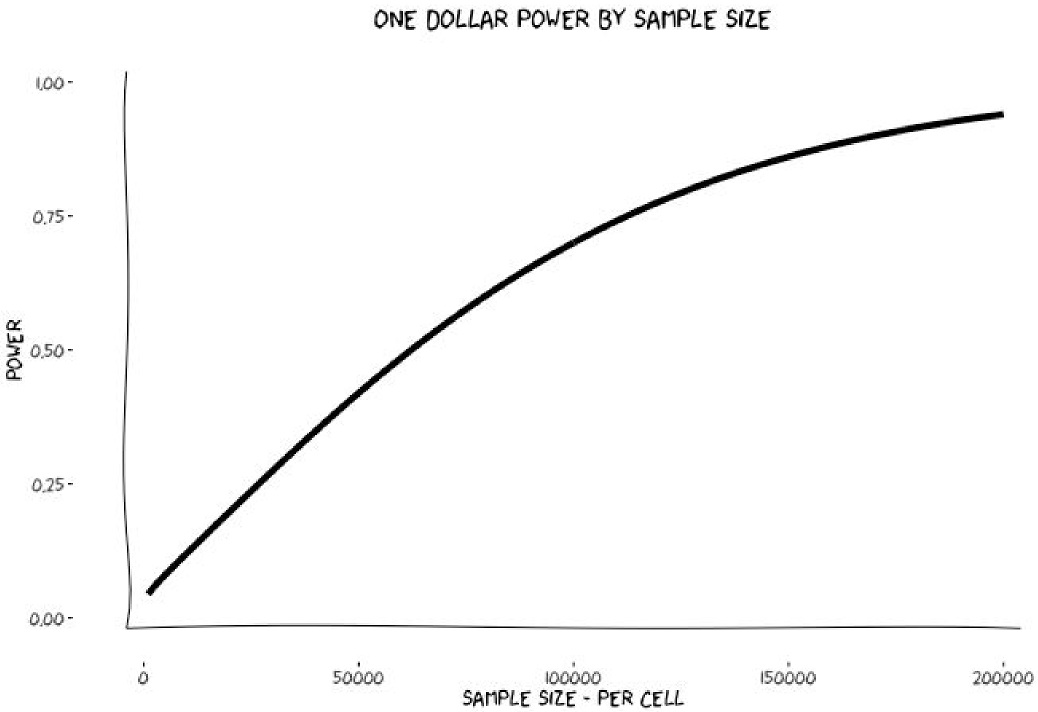
You need enough data to find your effect!

Power

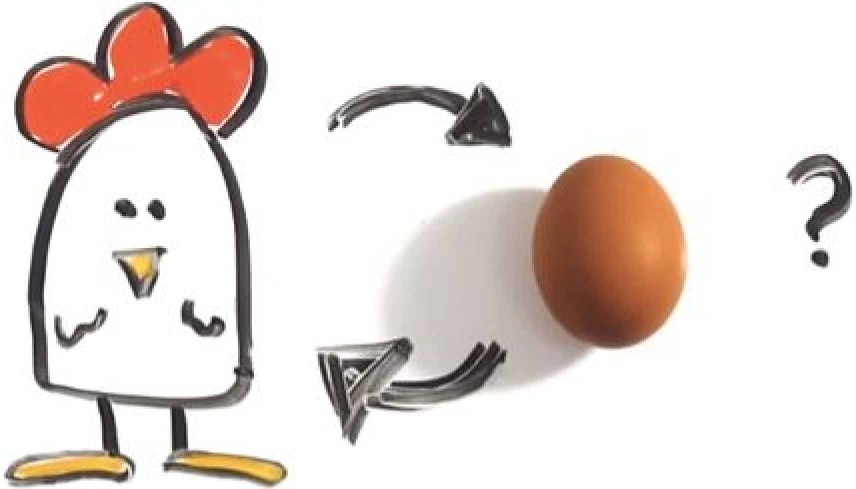
This is often called statistical “power”

POWER = 1 - P(Type II error)
= probability of detecting an effect when there is one

Example: detecting a $1 change in AOV



Often, the whole point of an experiment is to learn an effect size



What effect size are you looking for?

Often, the whole point of an experiment is to learn an effect size

- Prior domain knowledge - how big are effects?
- How small of an effect are you interested in?

Tricky. Starts to feel very “Bayesian”

Lesson 9: Lack of power causes false discoveries

False discovery rate

If you look back on your decisions, what percentage of "significant" findings will turn about to have been true?

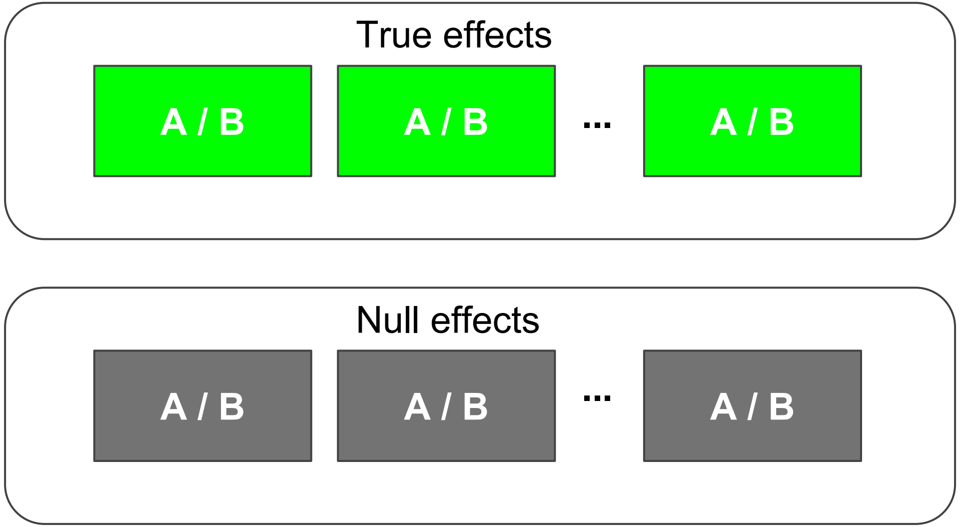
False discovery rate

If you look back on your decisions, what percentage of “significant” findings will turn about to have been true?

Beware your intuition - this depends on more than the probability of a Type I error!

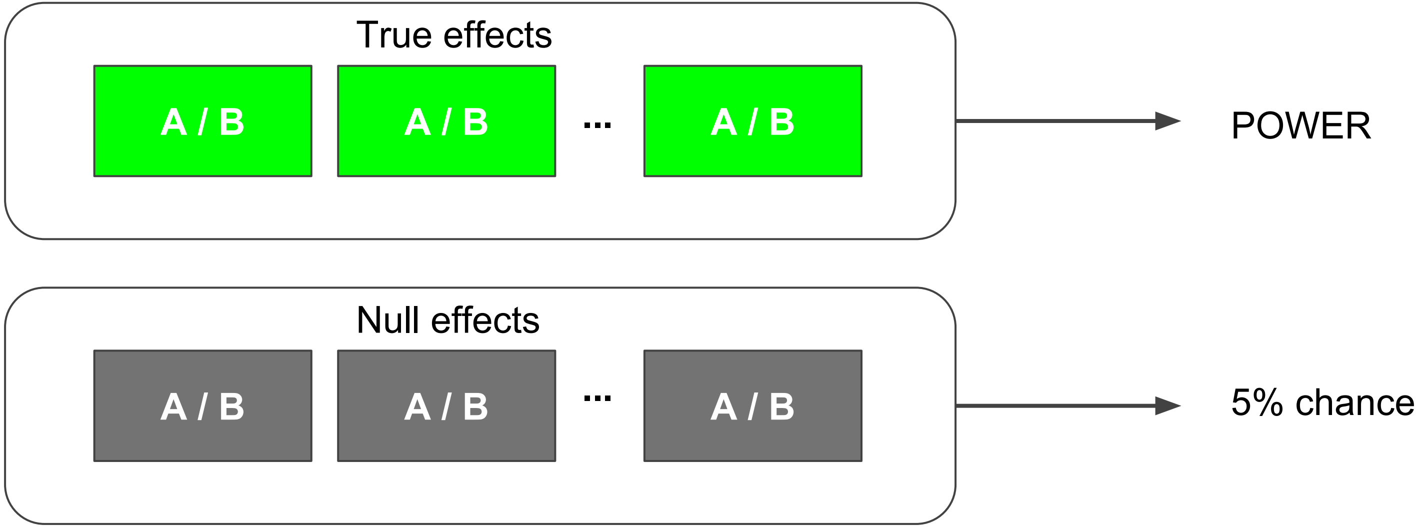
False discovery rate

What is the probability of claiming a significant effect?



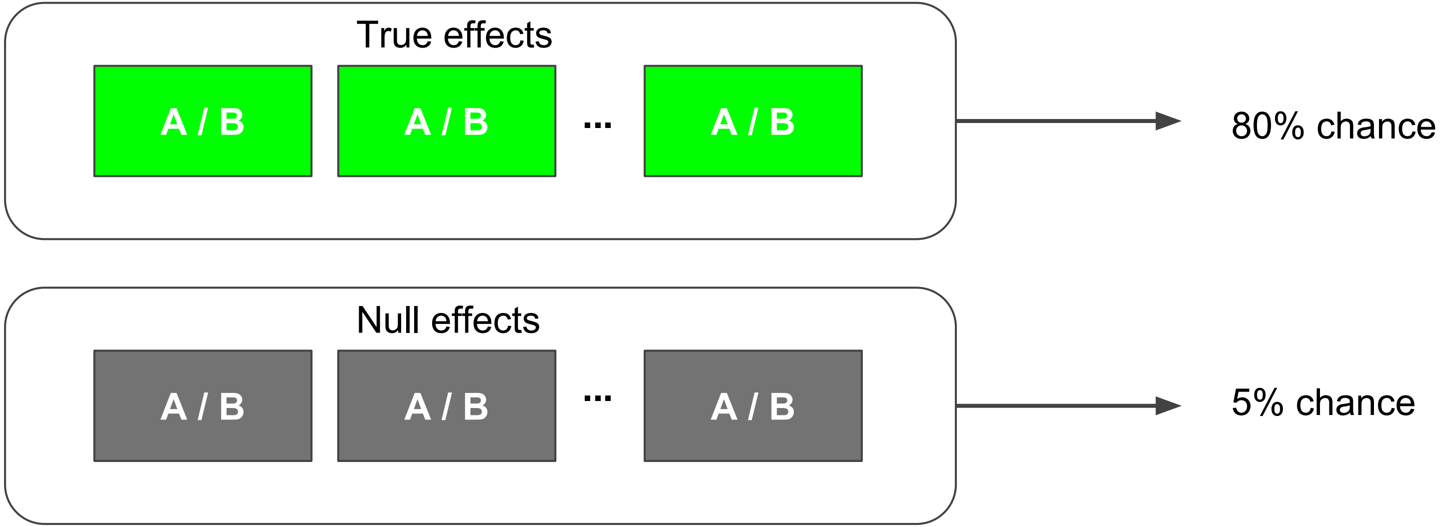
False discovery rate

What is the probability of claiming a significant effect?



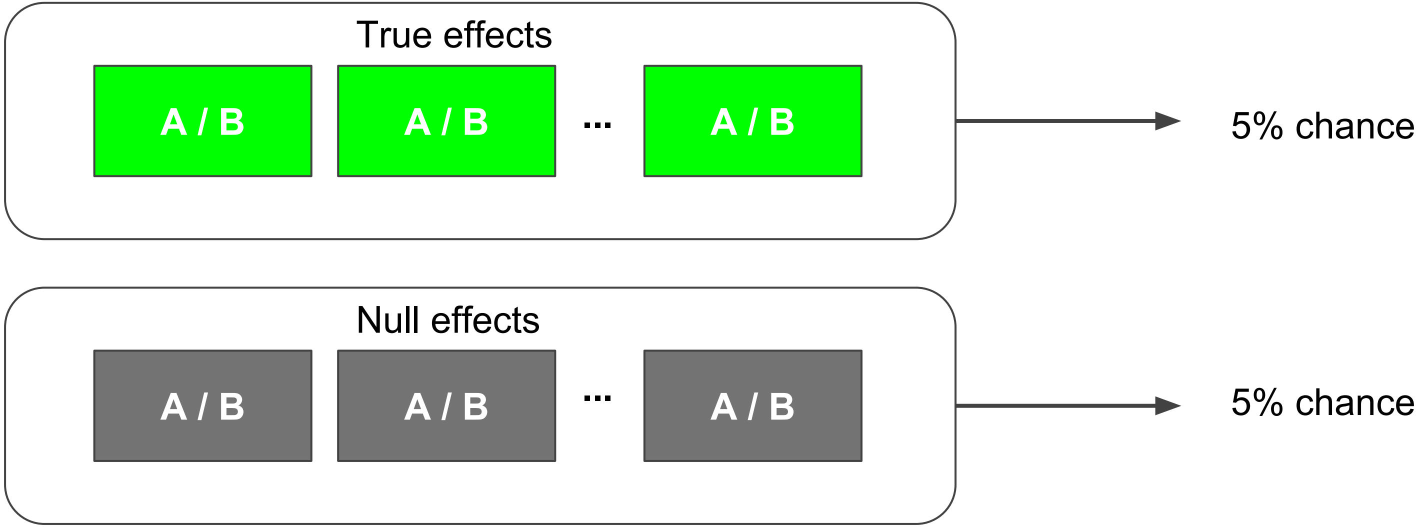
False discovery rate

We’d like power to be high



False discovery rate

But what if it is low?



False discovery rate

Given a significant result, what is the probability that it is actually a real effect?

False discovery rate

If you test an equal mix of true effects and null effects* the probability that a “significant” effect is actually real is

$$\frac{1}{1 + \frac{0.05}{\text{power}}}$$

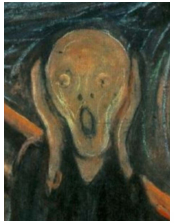
*big assumption

False discovery rate

If you test an equal mix of true effects and null effects the probability that a "significant" effect is actually real is

$$\frac{1}{1 + \frac{0.05}{\text{power}}}$$

- power = 0.05 -> prob your claim is real = 0.5



- power = 0.80 -> prob your claim is real = 0.94

False discovery rate

Running an underpowered experiment is often irresponsible!

Not just important at Stitch Fix

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there

http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

Lesson 10:
A/B testing isn’t magic

A/B testing isn't magic

At the end of the day, A/B is just a principled way to make comparisons

A/B testing isn't magic

At the end of the day, A/B is just a principled way to make comparisons

- It can't tell you what to optimize for
- It can't guarantee you'll find a globally optimal solution

A/B testing isn't magic

At the end of the day, A/B is just a principled way to make comparisons

- It can't tell you what to optimize for
- It can't guarantee you'll find a globally optimal solution

Key questions to take with you:

- Am I making a fair comparison (usually comes from randomization)
- What is my standard of evidence for this decision? Does the data meet it?

Appendix:
Bonus lessons

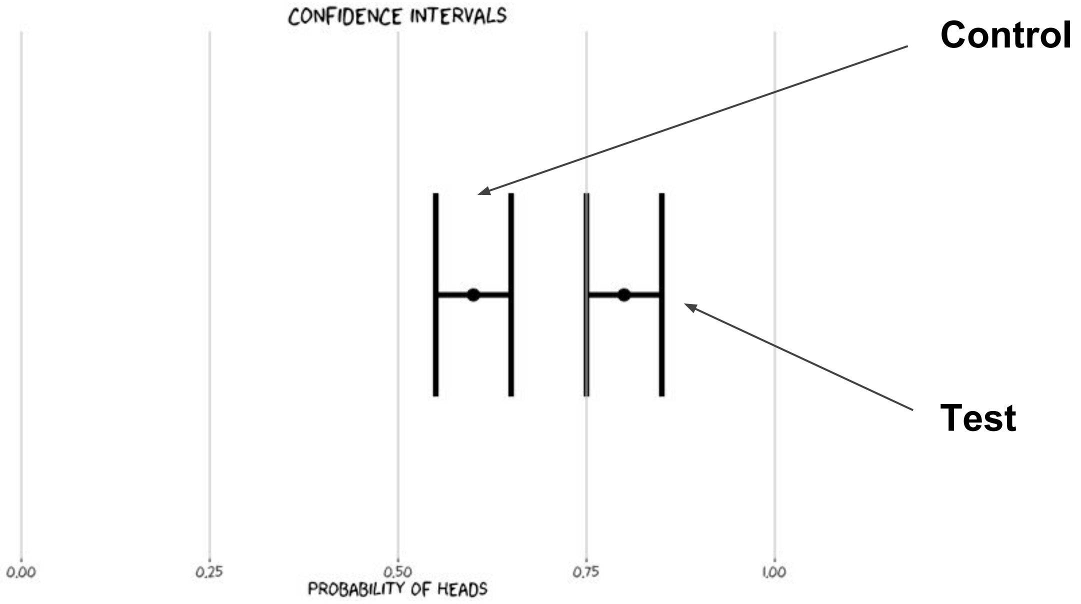
Lesson 11:
Differences deserve their own confidence interval

Comparing two groups

Often the results for two groups are shown as confidence intervals in the same figure

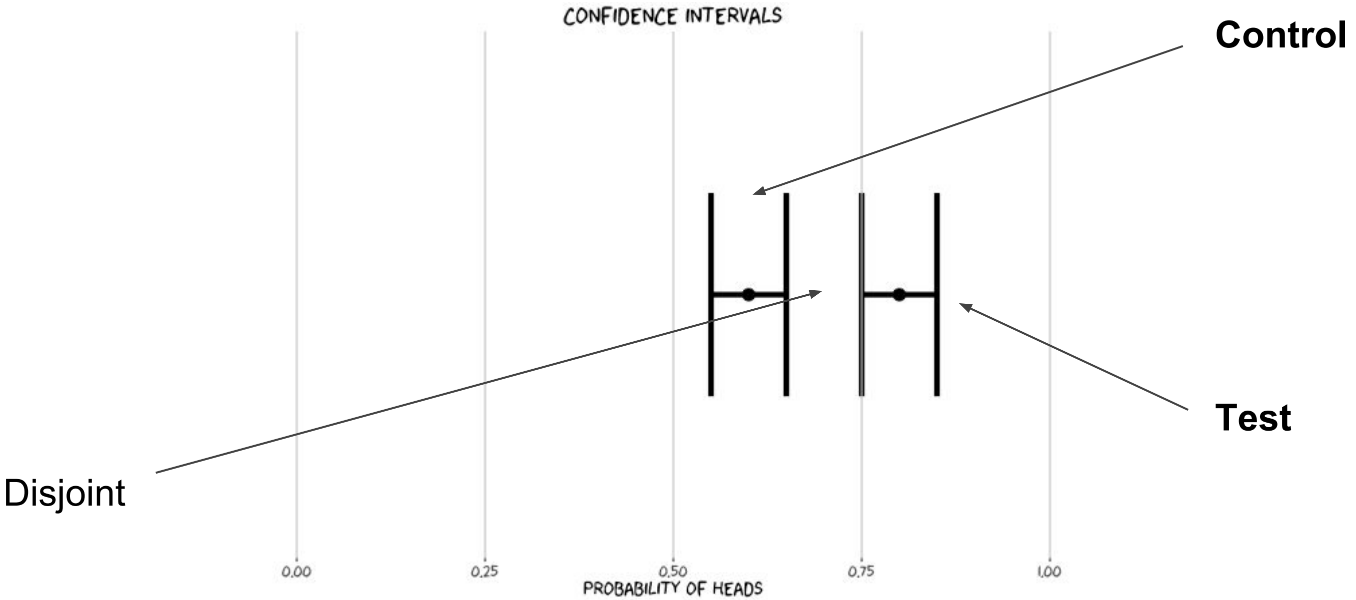
Comparing two groups

Often the results for two groups are shown as confidence intervals in the same figure



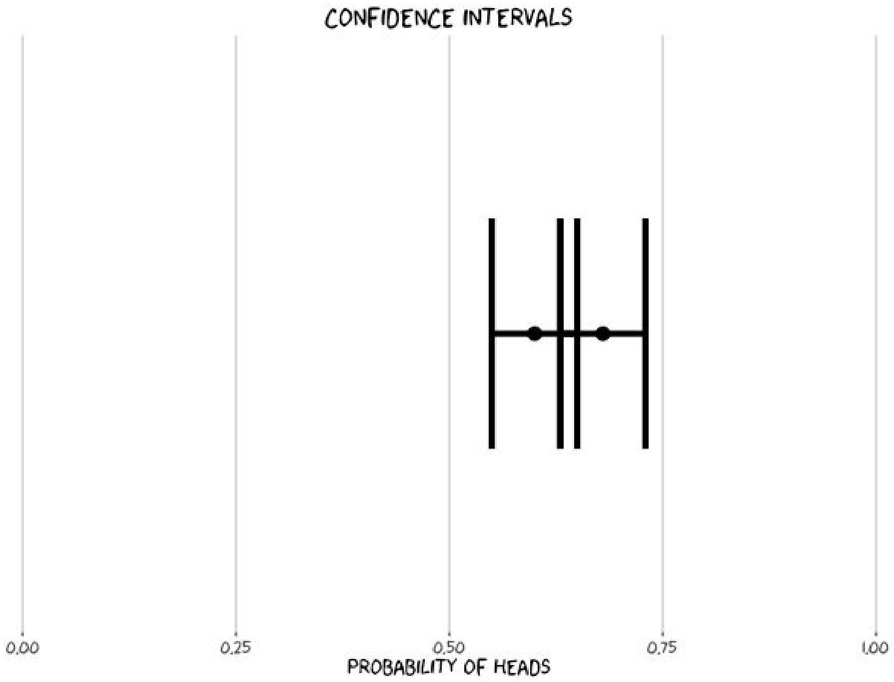
Disjoint intervals

If the intervals are disjoint, you can conclude the difference is significant!



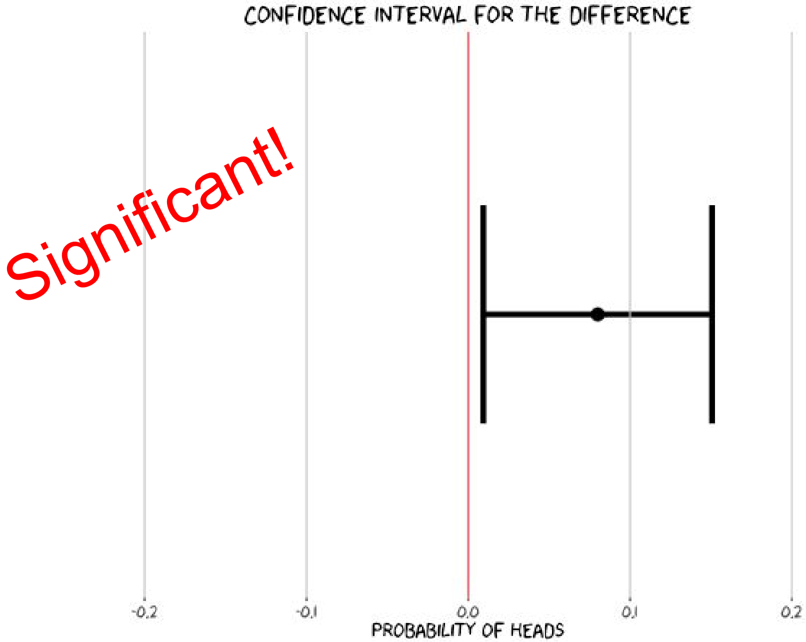
Overlapping intervals

But what if the intervals overlap? You can not *necessarily* conclude that difference is not significant



Confidence interval for the difference

To tell if the *difference* is significant you could look at the confidence interval for the difference and whether it contains zero!



Why? The confidence interval for a difference is narrower than sum of the width of the two confidence intervals

(standard deviation is subadditive for independent random variables)

Lesson 12:
Simultaneous experiments can create interference

Simultaneous experiments

It is not unusual for more than A/B test to be running at once

Is this a problem?

Simultaneous experiments

It is not unusual for more than A/B test to be running at once

Is this a problem?

Not usually, as long as each experiment is allocated randomly and the experiments do not *interact* in a meaningful way

[More info on simultaneous testing]

Example of non-malicious interference

Example: Styling Algos is running a (stylist-based) A/B test for the expanded RBM pilot

Example of non-malicious interference

Example: Styling Algos is running a (stylist-based) A/B test for the expanded RBM pilot

Suppose Team X wants to run a new test on *Bad Idea* - but is only interested in the impact on RBM.

Example of non-malicious interference

Example: Styling Algos is running a (stylist-based) A/B test for the expanded RBM pilot

Suppose Team X wants to run a new test on *Bad Idea* - but is only interested in the impact on RBM.

No problem! They can just look up the stylists using RBM and exclude all others from their experiment.

Example of non-malicious interference

Example: Styling Algos is running a (stylist-based) A/B test for the expanded RBM pilot

Suppose Team X wants to run a new test on *Bad Idea* - but is only interested in the impact on RBM.

No problem! They can just look up the stylists using RBM and exclude all others from their experiment.

But since *Bad Idea* is bad, the net impact will lower the performance of the RBM cell of the Styling Algo experiment!

Simultaneous experiments

Simultaneous experiments can interact.

Watch out for other changes/experiments that have an unequal impacts on your test or control cells

[More info on simultaneous testing]