Practical Testing

Nick Ross

UNIVERSITY OF
SAN FRANCISCO

October 15, 2017

Table of Contents

Disclaimer

NueBev is a fabricated company. TinyCo is a real company. Both of the cases described in this presentation are “not-real." They are amalgamations of a few different tests and situations that actually happened, repackaged in a way to highlight the underlying tensions and issues.

Who I am


UNIVERSITY OF
SAN FRANCISCO


SEGA®


TinyCo™


BATES
WHITE
ECONOMIC CONSULTING


THIRD FIN
CONSULTING

Practical Testing

- AB-testing and other experiments are frequently used by technology companies.
- We often assume that AB-testing is an “easy" type of experiment.
- This talk will cover:
  - Two “real-world" examples of difficult AB-tests.
  - Four common problems when running AB-tests.

Background

- Statistical experimentation frequently occurs on consumer facing products.
- Companies:
  - Have significant domain knowledge (unsurprisingly) which leads to (surprisingly) sophisticated knowledge.
  - Statistically insecure.
  - Mathematically inclined.
  - Not Google-sized, can't rely on scale to smooth-over errors.
- Mental model for firms attempting experimentation:
  - **Ed**, Engineering lead who is currently doing "data science."
  - **Cassie**, CEO who thinks the company could be more "data-driven".

Example #1: NueBev

- NueBev[1] provides subscription food / beverage.
- They charge a (hidden) margin (5%) per order.
- Customer specifies that they want 5 meals & beverages. They will pay $10.00 per person.
- Example: A once-a-month seminar / meeting.
- Company handles delivery and item selection.
- NueBev wishes to optimize this margin.
- Trade-off: If they set the margin too high, customers will quit.

Let's hear what Ed and Cassie say!

NueBev Analysis

- **Ed**: We need to run a test, so:
  - Lets test 5%, 7.5% and 10% margins.
  - Each month we have about 500 subscribers, so if we run the test for 6 months we will have 3,000 observations.
  - Our normal un-subscribe rate is 12% per-month, so, according to this online test calculator, we should get significance at the 5% level.
- **Cassie**: Why those margins? Why not 5%, 10% and 15%? Higher margins mean higher profits!

Failure

Reasons this test will fail:

1. Observation unit: The number of observations is closer to 500, not 500 $\times$ # months.
2. Unknown estimated effect: Does moving the margin from 5% to 7.5% actually change the customer experience?

Fixes

Simple fixes:

1. Observation unit is a single customer.
2. Use past data to estimate the “**customer-facing"** effect of moving to each new margin:
   - How much would the order change when moving to each margin level?
3. Modify the length of time of the test to insure that there is a measurable effects.

Example #2: TinyCo

- The company makes a Zoo game (“TinyZoo"). Users buy animals to put in their zoo.
- Company wants to test how much a particular item (Bannaminal) should cost: ($.99, $1.99 or $3.99)
- Decide to run an AB-test.
- This is a mobile (phone and iPad) game.

TinyZoo


Over 800 different animals to collect!
Tiny Zoo
6,999 XP
1000
ADD
1000
ADD
STORE
CUBBIES
LAB
BREEDING
COLLECTIONS
TRIENDS
TinyCo
New animals added daily!

TinyZoo Analysis

Ed: Let's run a 3 month AB-test from Nov-Jan:

- We have run these 3-month tests in the past and they work.
- Each *device* will receive one of three treatments ($.99, $1.99 or $3.99)
- After three months we will see which group had the highest revenue

Cassie: Sounds good!

Failure

- During the last week of December, the number of test subjects explodes 150% week-over-week.
- Why? *Device*
- Christmas – many users get new devices!
- Users ended up seeing multiple variants and complain.
- TinyCo ends up giving away the Banannimal for free.

Fixes

In this case, there was no “simple" fix.

General Failure

- Now that we have a sense of two AB-tests, let's talk about failure in a more general way.

Problem #1: Identification

- Common failure #1: Identification
- Identity classification:
  - Hardware / Software / Required Logins
  - Different combinations of the above result in Hybrid systems.
- *None of the above systems defines a user.*
- Users experience multiple, inconsistent, treatments.

Identification (cont.)

- Example:
  - 3% of users in treatment A
  - 4% of users in treatment B
  - ID error rate is 5%
- **Ed**: The difference between groups is 1%, but our user ID error rate is 5%. (Does this matter?)
- **Cassie**: Why do we even do these test? All I hear is caveats!

Problem #2: Social Effects

- Users talk to each other (IRL, Facebook)
  - One solution is to fence based on geography or language
  - Can results be generalized?
- **Ed**: We ran the test, users in Canada experienced treatment A, while users in the United States experienced treatment B. The p-value on the test shows that treatment B is statistically different than treatment A.
- **Cassie**: Do you really expect me to run this business based on Canada to United States comparisons?

Problem #3: Multiple KPI Problem

- Most products have more than one KPI. How should they be compared?
- **Ed**: We ran a test. Per-user engagement went up, but it looks like long-term retention went down.
- **Cassie**: So people are using our product more, but for a shorter time? Is that good? bad?

Problem #4: Tools and Knowledge

- Ed and Cassie are not statistically knowledgeable.
- Rely on tools that they do not fully understand.

Problem #4: Tools and Knowledge (cont.)

How Optimizely (Almost) Got Me Fired

Update: The folks at Optimizely let us know that they've launched a new statistical approach to address the concerns raised in this post.

It had been six months since we started concerted A/B testing efforts at SumAll, and we had come to an uncomfortable conclusion: most of our winning results were not translating into improved user acquisition. If anything, we were going sideways, and given that one of my chief responsibilities was to move the user acquisition needle, this was decidedly not good. Not good for me. Not good for my career. And not good for SumAll.

Conclusion

- Running an AB-test is difficult to get right and (hopefully!) this gives you a taste of why.

Conclusion

Questions?

Contact me at ncross@usfca.edu