

Online Experimentation at Microsoft

Ronny Kohavi, Experimentation Platform, Microsoft

Based on papers co-authored with Thomas Crook, Brian Frasca, and Roger Longbotham

<http://exp-platform.com/expMicrosoft.aspx>

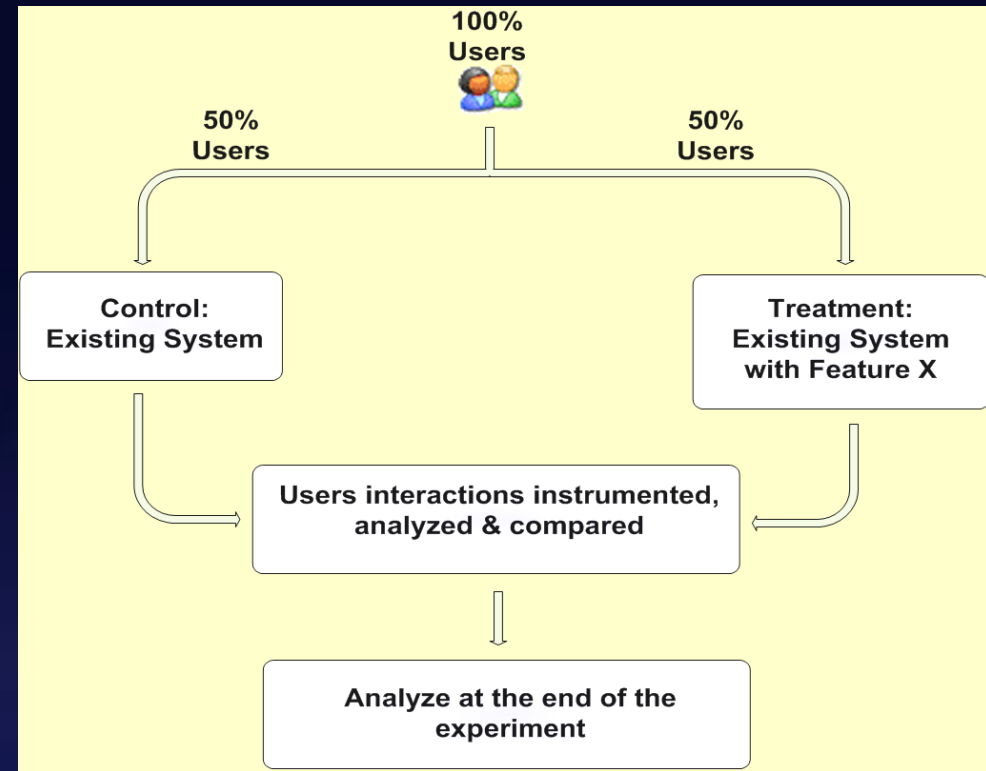
<http://exp-platform.com/ExpPpitfalls.aspx>

Goal / Problem Definition

- Goal: Accelerate software innovation through trustworthy experimentation
 - Enable a more scientific approach to planning and prioritization of features and designs
 - Trust is key: getting a number is easy. Getting a number you should trust is much harder.
The analyst may not realize problems.
- Experimentation is not applicable everywhere
 - Not covered in this talk: four necessary ingredients for experimentation to be useful (in paper, Section 6)
 - Sweet spot: websites and services

Controlled Experiments in One Slide

- Concept is trivial
 - Randomly split traffic between two (or more) versions
 - A/Control
 - B/Treatment
 - Collect metrics of interest
 - Analyze
- Best scientific way to prove **causality**, i.e., the changes in metrics are *caused* by changes introduced in treatment
- Must run statistical tests to confirm differences are not due to chance



Examples

- Three experiments that ran at Microsoft recently
- All had enough users for statistical validity
- Game: see how many you get right
 - Everyone please stand up
 - Three choices are:
 - A wins (the difference is statistically significant)
 - A and B are approximately the same (no stat sig diff)
 - B wins
 - If you guess randomly
 - $\frac{1}{3}$ left standing after first question
 - $\frac{1}{9}$ after the second question

MSN Real Estate

- “Find a house” widget variations
- Overall Evaluation Criterion: Revenue to Microsoft generated every time a user clicks search/find button

Find Your Dream Home or Apartment

City, State or ZIP

Existing homes New construction
 Foreclosures Rentals

Search listings ▶

A

Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale

Enter City State ▼

or

Enter Zip

Find homes ▶

B

- Raise your right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if you think they're about the same

MSN Real Estate

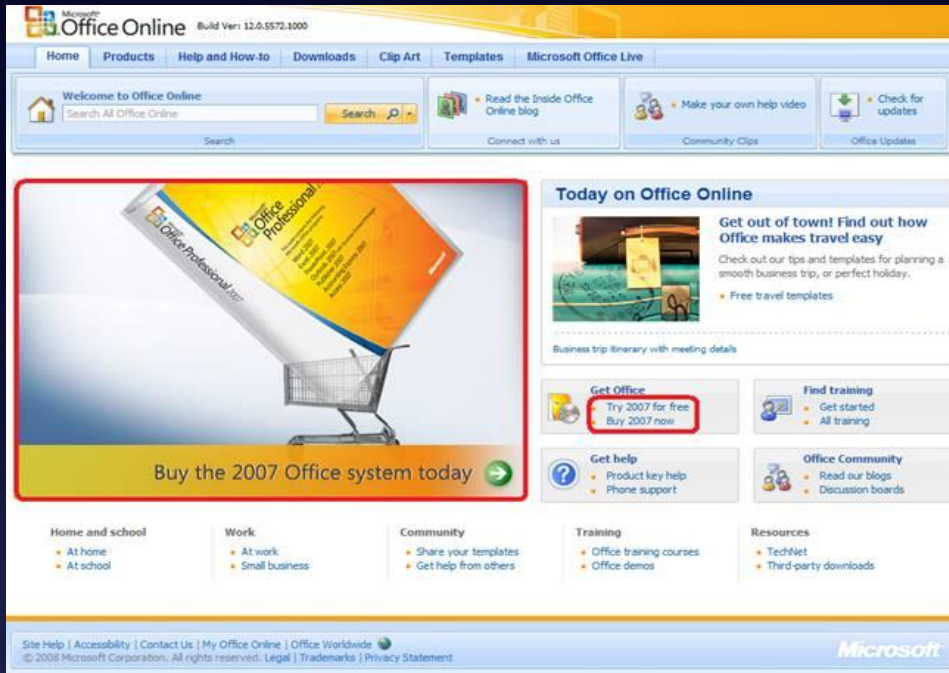
- If you did not raise a hand, please sit down
- If you raised your left hand, please sit down
- A was 8.5% better

Office Online

Test new design for Office Online homepage

OEC: Clicks on revenue generating links (red below)

A



B



- Raise your right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if you think they're about the same

Office Online

- If you did not raise a hand, please sit down
- If you raised your left hand, please sit down
- B was 64% worse

The Office Online team wrote

A/B testing is a fundamental and critical Web services... consistent use of A/B testing could save the company millions of dollars



MSN Home Page Search Box

OEC: Clickthrough rate for Search box and popular searches

A



Web | MSN | Images | Video | News | Maps | Shopping

  Live Search

Popular Searches: [Fireworks safety](#) | [Rihanna](#) | [Campaign patriotism](#)

B



Web | MSN | Images | Video | News | Maps | Shopping

 Live Search

[Fireworks safety](#) | [Rihanna](#) | [Campaign patriotism](#)

Differences: A has taller search box (overall size is the same), has magnifying glass icon, “popular searches”

B has big search button

- Raise your right hand if you think A Wins
- Raise your left hand if you think B Wins
- Don't raise your hand if they are the about the same

Search Box

- If you raised any hand, please sit down
- Insight
 - Stop debating, it's easier to get the data

US Search Box for Bing Launch

- For the launch of Bing, Microsoft's search engine, there was an effort to improve the search box on the MSN home page
- Control:



- Treatment:



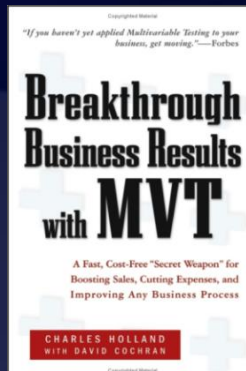
- New version was statistically significantly better
- Small changes have big impact

ROI on Experiments

- Estimated annual impact from multiple experiments (see paper for many examples) was over \$1M each
- How much of the value is due to the experimentation culture vs. the team having great ideas?
If the team just launched all ideas without testing, would they do well? ... Not even close!
- Small changes can have big impact, and large efforts sometimes had no impact or have negative impact
- Key observations: we are poor at predicting the value of ideas, and hence the criticality of getting actual data

Hard to Assess the Value of Ideas: Data Trumps Intuition

- Based on experiments with ExP at Microsoft
 - 1/3 of ideas were positive ideas and statistically significant
 - 1/3 of ideas were flat: no statistically significant difference
 - 1/3 of ideas were negative and statistically significant
- Our intuition is poor: 2/3rd of ideas do not improve the metric(s) they were designed to improve. Humbling!
- At Amazon, half of the experiments failed to show improvement
- QualPro tested 150,000 ideas over 22 years
 - 75 percent of important business decisions and business improvement ideas either have no impact on performance or actually hurt performance...



Key Lessons

- Avoid the temptation to try and build optimal features through extensive planning without early testing of ideas
- Experiment often
 - *To have a great idea, have a lot of them -- Thomas Edison*
 - *If you have to kiss a lot of frogs to find a prince, find more frogs and kiss them faster and faster -- Mike Moran, Do it Wrong Quickly*
- Try radical ideas. You may be surprised
 - Doubly true if it's cheap to implement (e.g., shopping cart recommendations and Behavior-Based search at Amazon)
 - *If you're not prepared to be wrong, you'll never come up with anything original – [Sir Ken Robinson](#), TED 2006*


Cultural Challenges

- Adaptive changes are hard. Microsoft knows how to ship shrink-wrapped software; less experience in online world
- Education is important
 - We started teaching a monthly $\frac{3}{4}$ -day class on experimentation
 - Initially, we couldn't fill it; now wait listed; > 500 people attended
- Published papers in KDD to establish credibility inside Microsoft, and get reviewer feedback, which was highly beneficial
- We use the HiPPO as our mascot and give stress HiPPOs at talks with our URL. Acronym: Highest Paid Person's Opinion



Unique Posters to Raise Awareness

- We put posters across the campus to raise awareness
- Usually in pairs with unique URLs to A/B test them
- And in weird places...
- Experiment or Die won the first round



Experiment or Die!

Do you want to transform the way we build online services at Microsoft? Do you want to have impact across multiple products and services, such as Live, MSN and Microsoft.com? Then join the Microsoft Experimentation Platform team.

The Experimentation Platform is now hiring SDEs, SDETs and PMs
To learn more about Experimentation Platform job opportunities, visit <http://experiment/jobs>

Our mission is to accelerate software innovation through trustworthy experimentation



Some Controversy is Fine

- A director of communications complained that the “Experiment or Die” poster is “threatening.” We explained the usage
- V2 was our most successful poster
 - Great image
 - Quotation from Google’s Hal Varian

Experiment or Die!

“Being able to figure out quickly what works and what doesn’t can mean the difference between survival and extinction.”

-- Hal Varian, Google Chief Economist

Visit <http://experiment/die> to learn more about our classes, job openings and intro lunch talks.



ExP • INNOVATE • TRUST

Microsoft

Our mission is to accelerate software innovation through trustworthy experimentation

The Cultural Challenge

It is difficult to get a man to understand something when his salary depends upon his not understanding it.

-- Upton Sinclair

- Why people/orgs avoid controlled experiments
 - Some believe it threatens their job as decision makers
 - At Microsoft, program managers select the next set of features to develop. Proposing several alternatives and admitting you don't know which is best is hard
 - Editors and designers get paid to select a great design
 - Failures of ideas may hurt image and professional standing. It's easier to declare success when the feature launches
 - We've heard: "we know what to do. It's in our DNA," and "why don't we just do the right thing?"

A Few Pitfalls

- Getting a number is easy.
Getting a number you should trust is much harder
- In the paper we shared seven pitfalls
- Here are three

Best Practice: Ramp-up



- Ramp-up
 - Start an experiment at 0.1%
 - Do some simple analyses to make sure no egregious problems can be detected
 - Ramp-up to a larger percentage, and repeat until 50%
- Big differences are easy to detect because the min sample size is quadratic in the effect we want to detect
 - Detecting 10% difference requires a small sample and serious problems can be detected during ramp-up
 - Detecting 0.1% requires a population $100^2 = 10,000$ times bigger
- Abort the experiment if treatment is significantly worse on OEC or other key metrics (e.g., time to generate page)

Pitfall : Combining Data During Ramp-up

- Simplified example: 1,000,000 users per day

Conversion Rate for two days	Friday	Saturday	Total
	C/T split: 99/1	C/T split: 50/50	
Control	$\frac{20,000}{990,000} = 2.02\%$	$\frac{5,000}{500,000} = 1.00\%$	$\frac{25,000}{1,490,000} = 1.68\%$
Treatment	$\frac{230}{10,000} = 2.30\%$	$\frac{6,000}{500,000} = 1.20\%$	$\frac{6,230}{510,000} = 1.22\%$

- For each individual day the Treatment is much better
- However, cumulative result for Treatment is **worse**
- This counter-intuitive effect is called Simpson's paradox

Best Practice: A/A Test

- Run A/A tests
 - Run an experiment where the Treatment and Control variants are coded identically and validate the following:
 1. Are users split according to the planned percentages?
 2. Is the data collected matching the system of record?
 3. Are the results showing non-significant results 95% of the time?

This is a powerful technique for finding bugs and other integration issues **before** teams try to make data-driven decisions

- Multiple integrations at Microsoft failed A/A tests
 - Example problem: caching issues. If variants share an LRU cache and Control is 90% while treatment is 10%, control pages will be cached more often and be faster

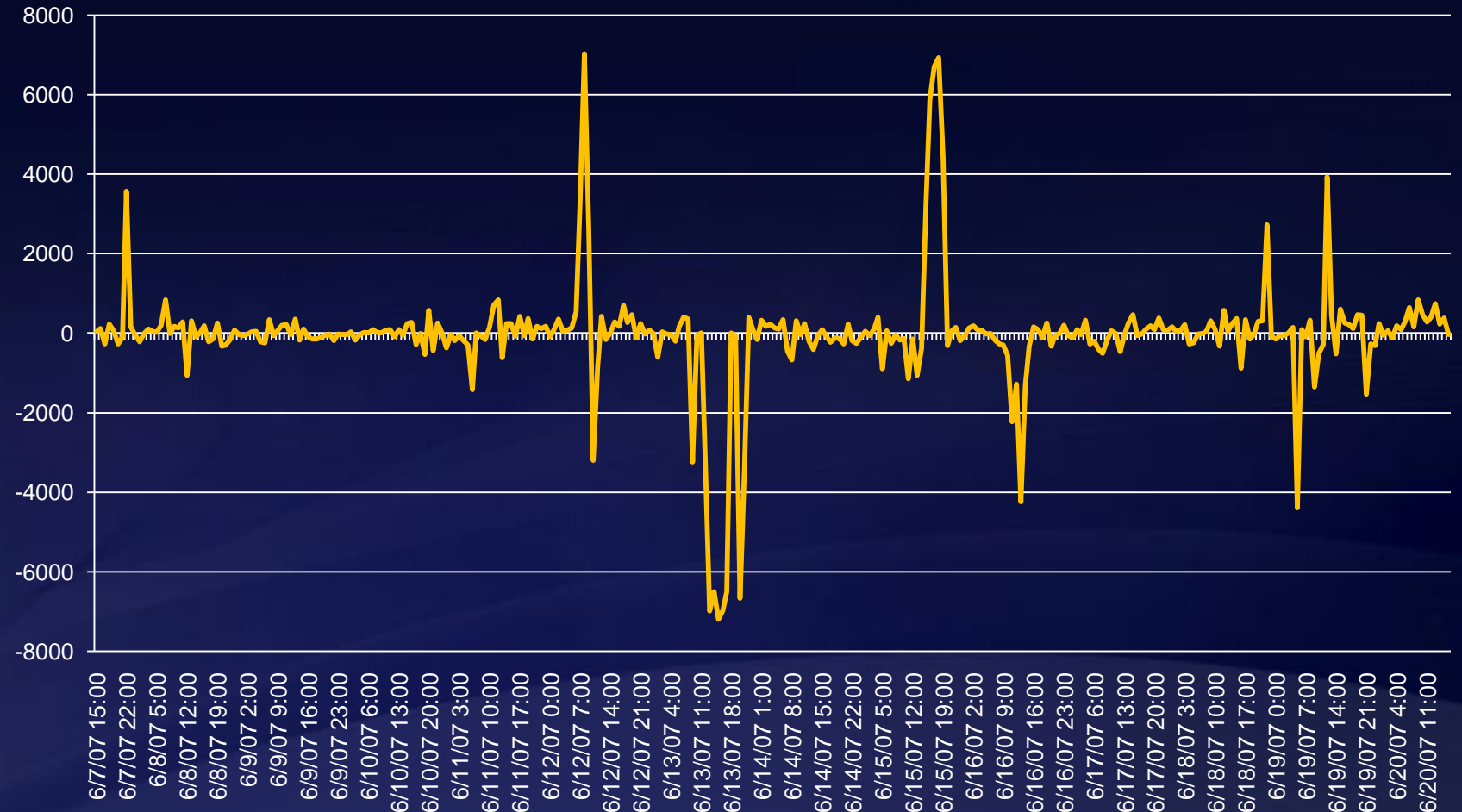
Pitfall: Not Filtering out Robots

- Internet sites can get a significant amount of robot traffic (search engine crawlers, email harvesters, botnets, etc.)
- Robots can cause misleading results
 - Most concerned about robots with high traffic (e.g. clicks or PVs) that stay in Treatment or Control (we've seen one robot with > 600,000 clicks in a month on one page)
- Identifying robots can be difficult
 - Some robots identify themselves
 - Many look like human users and even execute JavaScript
 - Use heuristics to ID and remove robots from analysis (e.g. more than 100 clicks in an hour)

Effect of Robots on A/A Experiment

- Each hour represents clicks from thousands of users
- The “spikes” can be traced to single “users” (robots)

Clicks for Treatment minus Control by Hour for A/A test
No Robots Removed



The OEC

- OEC = Overall Evaluation Criterion
 - Agree early on what you are optimizing
 - Getting agreement on the OEC in the org is a huge step forward
 - Suggestion: optimize for customer lifetime value, not immediate short-term revenue
 - Criterion could be weighted sum of factors, such as
 - Time on site (per time period, say week or month)
 - Visit frequency
 - Report many other metrics for diagnostics, i.e., to understand the why the OEC changed and raise new hypotheses

OEC Thought Experiment

- Tiger Woods comes to you for advice on how to spend his time: improving golf, or improving ad revenue (most revenue comes from ads)
 - Short term, he could improve his ad revenue by focusing on ads...
 - But to optimize lifetime financial value (and immortality as a great golf player), he needs to focus on the game
 - While the example seems obvious, organizations commonly make the mistake of focusing on the short term



Pitfall: Wrong OEC

- In the Office Online example, the treatment had a drop in the OEC of 64%, but it was clickthrough
 - Were sales for Treatment correspondingly less also?
 - Our interpretation is that not having the price shown in the Control lead more people to click to determine the price
- Lesson: measure what you really need to measure, even if it's difficult!
- What's a good OEC for the MSN home page?
 - Click-through rate per user
 - Clicks per user (captures increased visit frequency)
 - (sum of (Click * estimated value-of-click)) per user

Summary

- Our goal is to accelerate software innovation through trustworthy experimentation
- We built an Experimentation Platform to reduce the costs of running experiments and decrease complexity of analysis
- Cultural and adaptive challenges are tough.
We addressed them through education, awareness (posters), publications (building credibility), and most important: successful experiments with significant ROI
- Trust is key: getting a number is easy.
Getting a number you should trust is much harder

<http://exp-platform.com>



Accelerating software Innovation through trustworthy experimentation

Appendix

Stress HiPPO

The less data, the stronger the opinions

- Whenever you feel stressed that a decision is made without data, squeeze the Stress-HiPPO
- Put one in your office to show others you believe in data-driven decisions based on experiments
- Hippos kill more humans than any other (non-human) mammal (really)
- Don't let HiPPOs in your org kill innovative ideas. ExPeriment!



Advantages of Controlled Experiments

- Controlled experiments test for **causal** relationships, not simply correlations
- When the variants run concurrently, only two things could explain a change in metrics:
 1. The “feature(s)” (A vs. B)
 2. Random chance

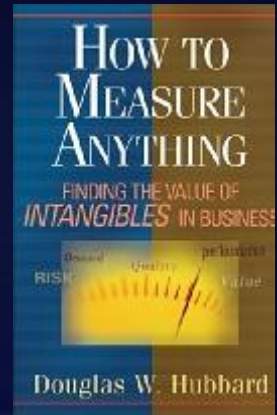
Everything else happening affects both the variants

For #2, we conduct statistical tests for significance
- The gold standard in science and the only way to prove efficacy of drugs in FDA drug tests

Issues with Controlled Experiments (1 of 2)

If you don't know where you are going, any road will take you there
—Lewis Carroll

- Org has to agree on OEC (Overall Evaluation Criterion).
 - This is hard, but it provides a clear direction and alignment
 - Some people claim their goals are “soft” or “intangible” and cannot be quantified. Think hard and read Hubbard’s *How to Measure Anything: Finding the Value of Intangibles in Business*
- Quantitative metrics, not always explanations of “why”
 - A treatment may lose because page-load time is slower. At Amazon, we slowed pages by 100-250msec and lost 1% of revenue
 - A treatment may have JavaScript that fails on certain browsers, causing users to abandon.



Issues with Controlled Experiments (2 of 2)

- Primacy/newness effect
 - Changing navigation in a website may degrade the customer experience (temporarily), even if the new navigation is better
 - Evaluation may need to focus on new users, or run for a long period
- Multiple experiments
 - Even though the methodology shields an experiment from other changes, statistical variance increases making it harder to get significant results. There can also be strong interactions (rarer than most people think)
- Consistency/contamination
 - On the web, assignment is usually cookie-based, but people may use multiple computers, erase cookies, etc. Typically a small issue
- Launch events / media announcements sometimes preclude controlled experiments
 - The journalists need to be shown the “new” version

Run Experiments at 50/50%

- Novice experimenters run 1% experiments
- To detect an effect, you need to expose a certain number of users to the treatment (based on power calculations)
- Fastest way to achieve that exposure is to run equal-probability variants (e.g., 50/50% for A/B)
- If you perceive risk, don't start an experiment at 50/50% from the beginning: Ramp-up over a short period

Randomization



- Good randomization is critical.
It's unbelievable what mistakes developers will make in favor of efficiency
- Properties of user assignment
 - Consistent assignment. User should see the same variant on successive visits
 - Independent assignment. Assignment to one experiment should have no effect on assignment to others (e.g., Eric Peterson's code in his book gets this wrong)
 - Monotonic ramp-up. As experiments are ramped-up to larger percentages, users who were exposed to treatments must stay in those treatments (population from control shifts)

Simpson's Paradox

- Lack of awareness can lead to mistaken conclusions about causality
- Unlike esoteric brain teasers, it happens in real life. My team at Blue Martini spent days debugging our software once, but it was fine
- In the next few slides I'll share examples that seem "impossible"
- We'll then explain why they are possible and do happen
- Discuss implications/warning

Examples 1: Drug Treatment

- Real-life example for kidney stone treatments
- Overall success rates:
 - Treatment A succeeded 78%, Treatment B succeeded 83% (better)
- Further analysis splits the population by stone size
 - For small stones
Treatment A succeeded 93% (better), Treatment B succeeded 87%
 - For large stones
Treatment A succeeded 73% (better), Treatment B succeeded 69%
 - Hence treatment A is better in both cases, yet was worse in total
- People going into treatment have either small stones or large stones
- A similar real-life example happened when the two populations segments were cities (A was better in each city, but worse overall)

Example 2: Sex Bias?

- Adopted from real data for UC Berkeley admissions
- Women claimed sexual discrimination
 - **Only 34% of women were accepted, while 44% of men were accepted**
- Segmenting by departments to isolate the bias, they found that **all** departments accept a higher percentage of women applicants than men applicants.
(If anything, there is a slight bias in favor of women!)
- There is no conflict in the above statements.
It's possible and it happened

Last Example: Batting Average

- Baseball example
 - One player can hit for a higher batting average than another player during the first half of the year
 - Do so again during the second half
 - But to have a lower batting average for the entire year

- Example

	First Half	Second Half	Total season
A	$4 / 10 = 0.400$	$25 / 100 = 0.250$	$29 / 110 = 0.264$
B	$35 / 100 = 0.350$	$2 / 10 = 0.200$	$37 / 110 = 0.336$

- Key to the “paradox” is that the segmenting variable (e.g., half year) interacts with “success” and with the counts.
E.g., “A” was sick and rarely played in the 1st half, then “B” was sick in the 2nd half, but the 1st half was “easier” overall.

Not Really a Paradox, Yet Non-Intuitive

- If $a/b < A/B$ and $c/d < C/D$, it's possible that $(a+c)/(b+d) > (A+C)/(B+D)$
- We are essentially dealing with weighted averages when we combine segments

Important, not Just a Cool Teaser

- Why is this so important?
- In knowledge discovery, we state probabilities (correlations) and associate them with causality
 - Treatment T1 works better
 - Berkeley discriminates against women
- We must be careful to check for confounding variables, which may be latent hidden
- With Controlled Experiments, we scientifically prove causality (But Simpson's paradox can, and does occur, if different days use different proportions for control/treatment.)