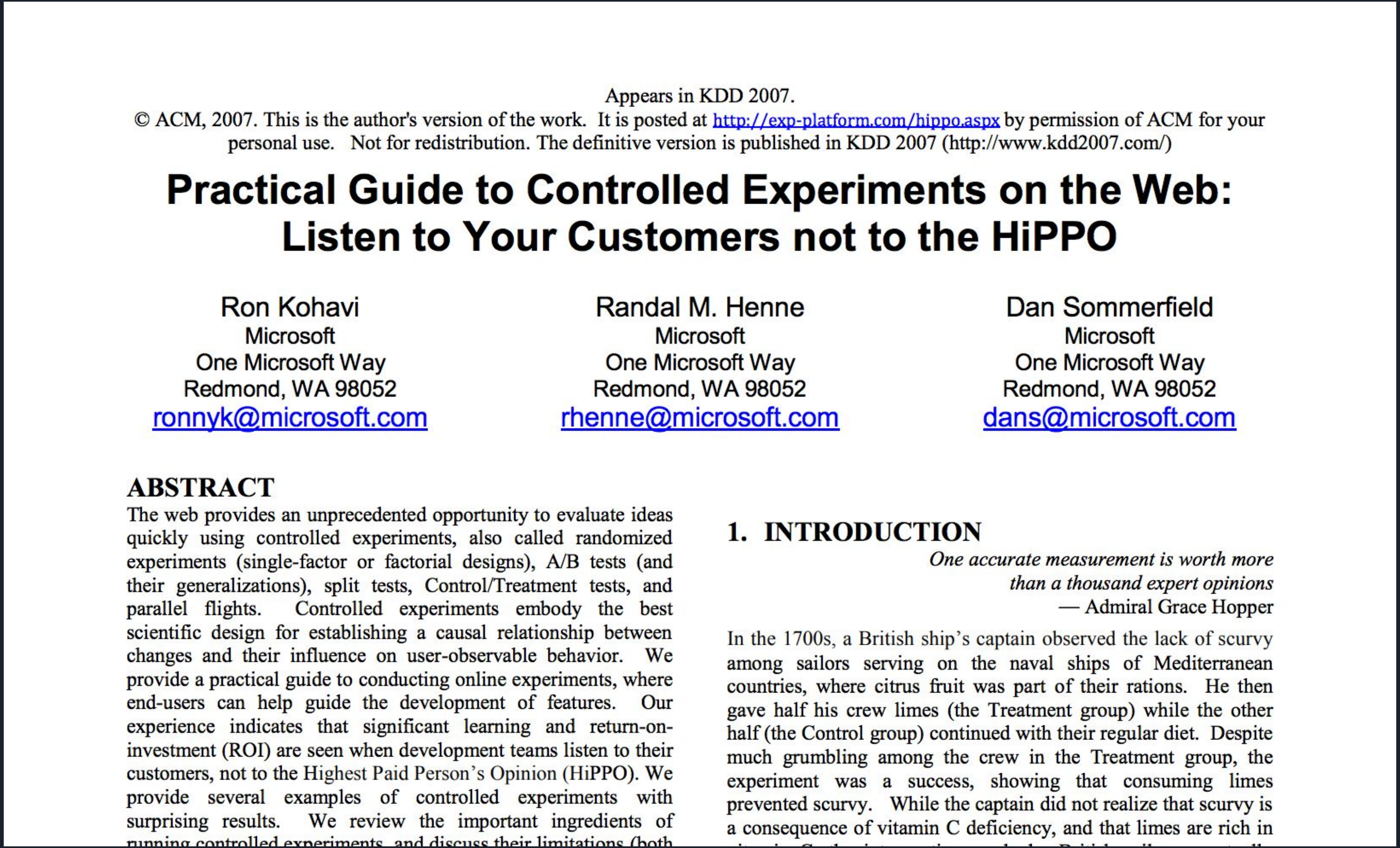
Online Experiments for Beginners

By: Neil Menne
Twitter: @the1evilgenius
Github: NeilMenne

Overview

- Motivation
- Constructing Our Hypothesis
- Understanding Our Users
- Analyzing/Interpreting the Results
- Lessons Learned

Personal Journey



Appears in KDD 2007.

Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO

Ron Kohavi
Microsoft
One Microsoft Way
Redmond, WA 98052
ronnyk@microsoft.com

Randal M. Henne
Microsoft
One Microsoft Way
Redmond, WA 98052
rhenne@microsoft.com

Dan Sommerfield
Microsoft
One Microsoft Way
Redmond, WA 98052
dans@microsoft.com

ABSTRACT

The web provides an unprecedented opportunity to evaluate ideas quickly using controlled experiments, also called randomized experiments (single-factor or factorial designs), A/B tests (and their generalizations), split tests, Control/Treatment tests, and parallel flights. Controlled experiments embody the best scientific design for establishing a causal relationship between changes and their influence on user-observable behavior. We provide a practical guide to conducting online experiments, where end-users can help guide the development of features. Our experience indicates that significant learning and return-on-investment (ROI) are seen when development teams listen to their customers, not to the Highest Paid Person's Opinion (HiPPO). We provide several examples of controlled experiments with surprising results. We review the important ingredients of running controlled experiments, and discuss their limitations (both

1. INTRODUCTION

One accurate measurement is worth more than a thousand expert opinions
— Admiral Grace Hopper

In the 1700s, a British ship's captain observed the lack of scurvy among sailors serving on the naval ships of Mediterranean countries, where citrus fruit was part of their rations. He then gave half his crew limes (the Treatment group) while the other half (the Control group) continued with their regular diet. Despite much grumbling among the crew in the Treatment group, the experiment was a success, showing that consuming limes prevented scurvy. While the captain did not realize that scurvy is a consequence of vitamin C deficiency, and that limes are rich in

Learn from the Best

- HiPPO: Highest Paid Person’s Opinion
- Understanding your users
- Retention

Constructing Our Hypothesis

Pardon My Terminology

- Control
- Treatment
- Null Hypothesis
- Unit of Experimentation
- Overall Evaluation Criterion

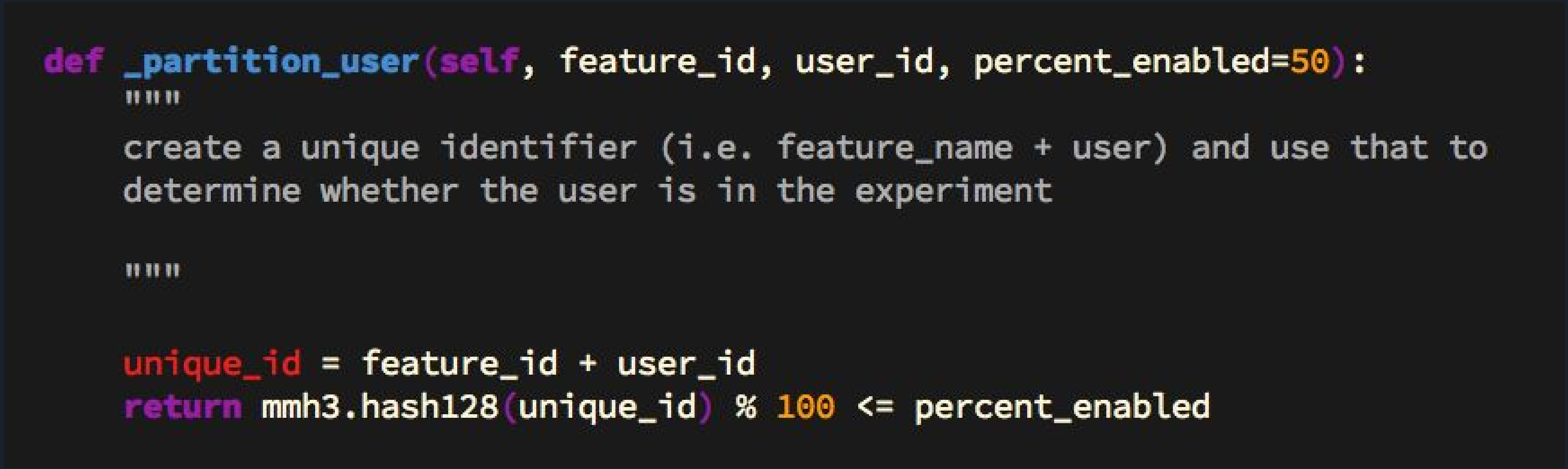
Measure This!

- Robust vs Sensitive
- Short-Term Measurable vs Long-Term Focused

Sawing the Users in Half

- Desirable Properties
  - Consistent User Experience
  - Independently Considered per Experiment
- Approaches
  - (Pseudo)-RNG
  - Hashing

teh codez



```python
def _partition_user(self, feature_id, user_id, percent_enabled=50):
    """
    create a unique identifier (i.e. feature_name + user) and use that to
    determine whether the user is in the experiment

    """

    unique_id = feature_id + user_id
    return mmh3.hash128(unique_id) % 100 <= percent_enabled
```

Stating our Hypothesis

- Null Hypothesis, H_0, is that there is no statistical difference in the control and treatment
- Alternate Hypothesis, H_1, there *is* a statistical difference

Understanding our Users

Check Yourself Into A/A

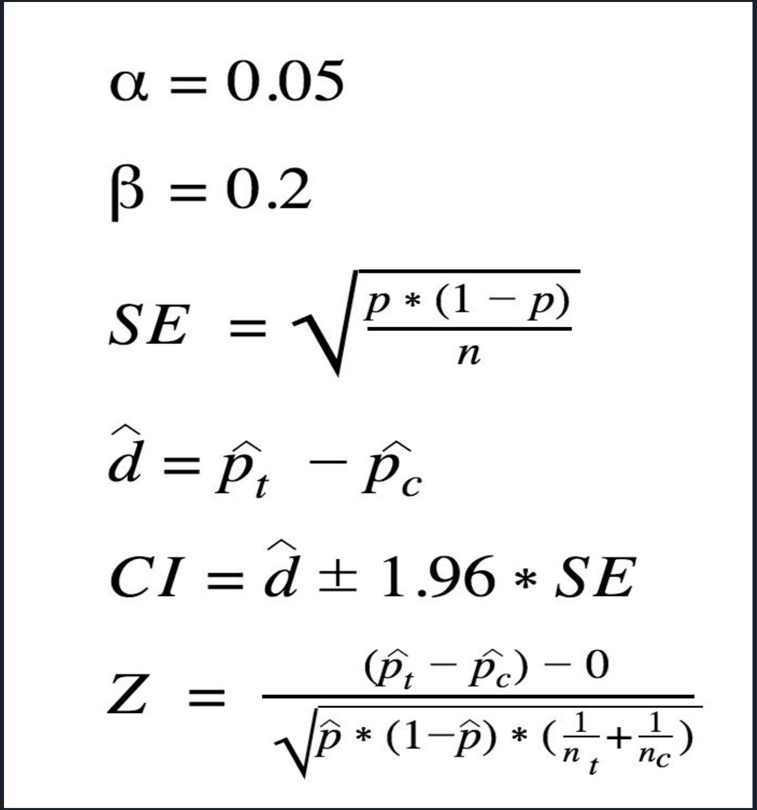
- Establish baselines for your metrics
- Understand the variability of your data
- Verify our instrumentation/partitioning

Initial A/A Results

	Control	Treatment (Just Control)
Converted:	2159	2107
Total Users:	4335	4213

Just enough math to see us through...

- Confidence Level
- Power
- Standard Error
- Difference
- Confidence Interval
- Z-Score



$$\alpha = 0.05$$

$$\beta = 0.2$$

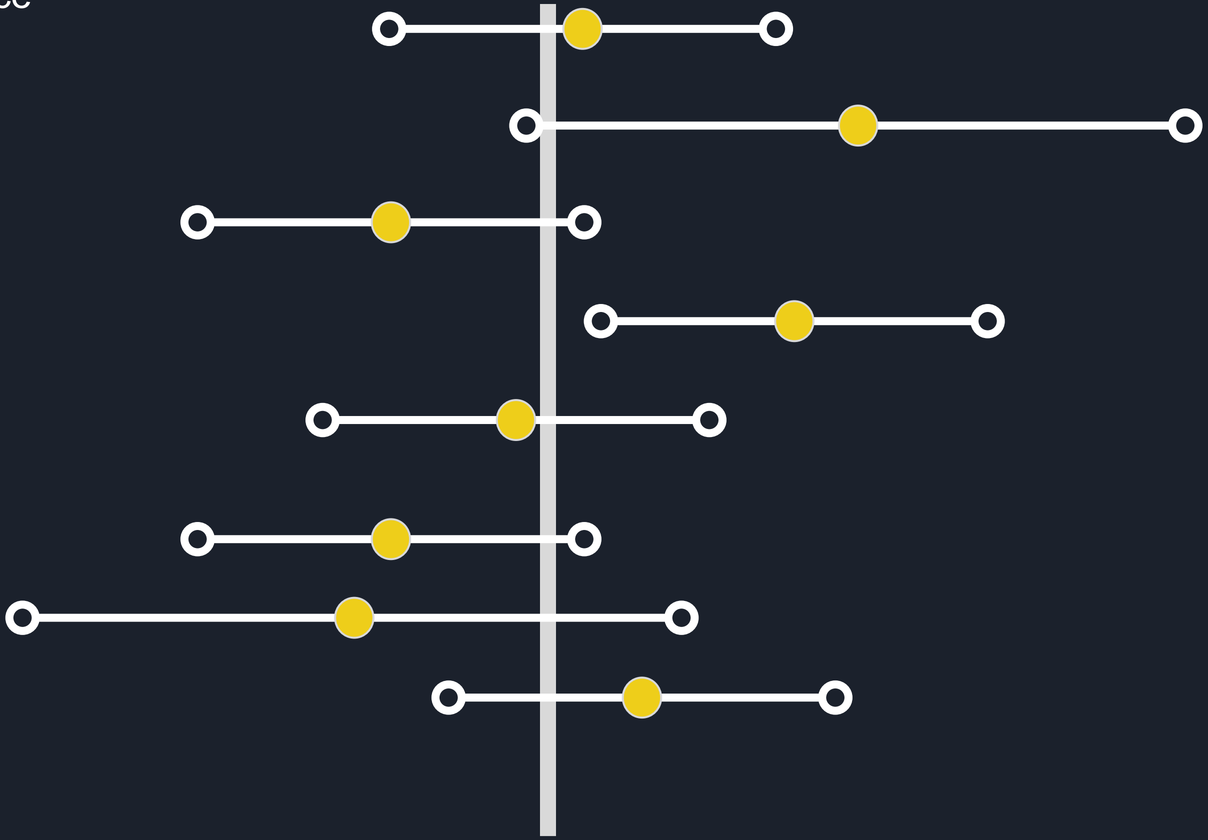
$$SE = \sqrt{\frac{p * (1 - p)}{n}}$$

$$\hat{d} = \hat{p}_t - \hat{p}_c$$

$$CI = \hat{d} \pm 1.96 * SE$$

$$Z = \frac{(\hat{p}_t - \hat{p}_c) - 0}{\sqrt{\hat{p} * (1 - \hat{p}) * (\frac{1}{n_t} + \frac{1}{n_c})}}$$

Visualizing Confidence



“True Value”

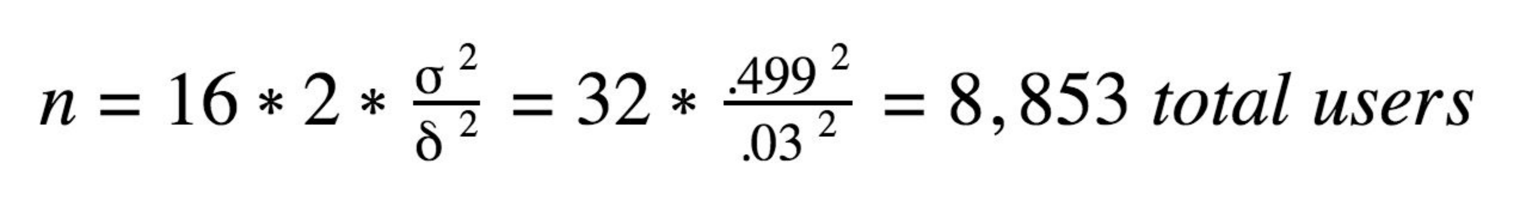
Analyzing Results

Solve for Z

Control Probability	49.80% (i.e. 2159/4335)
Treatment Probability	50.01% (i.e. 2107/4213)
Absolute Difference	0.21%
Test Statistic, Z	0.19
Standard Error	0.77%
Confidence Interval	(49.24%, 50.78%)

Sizing an Experiment

- Proportion of users per variant
- Duration of the experiment

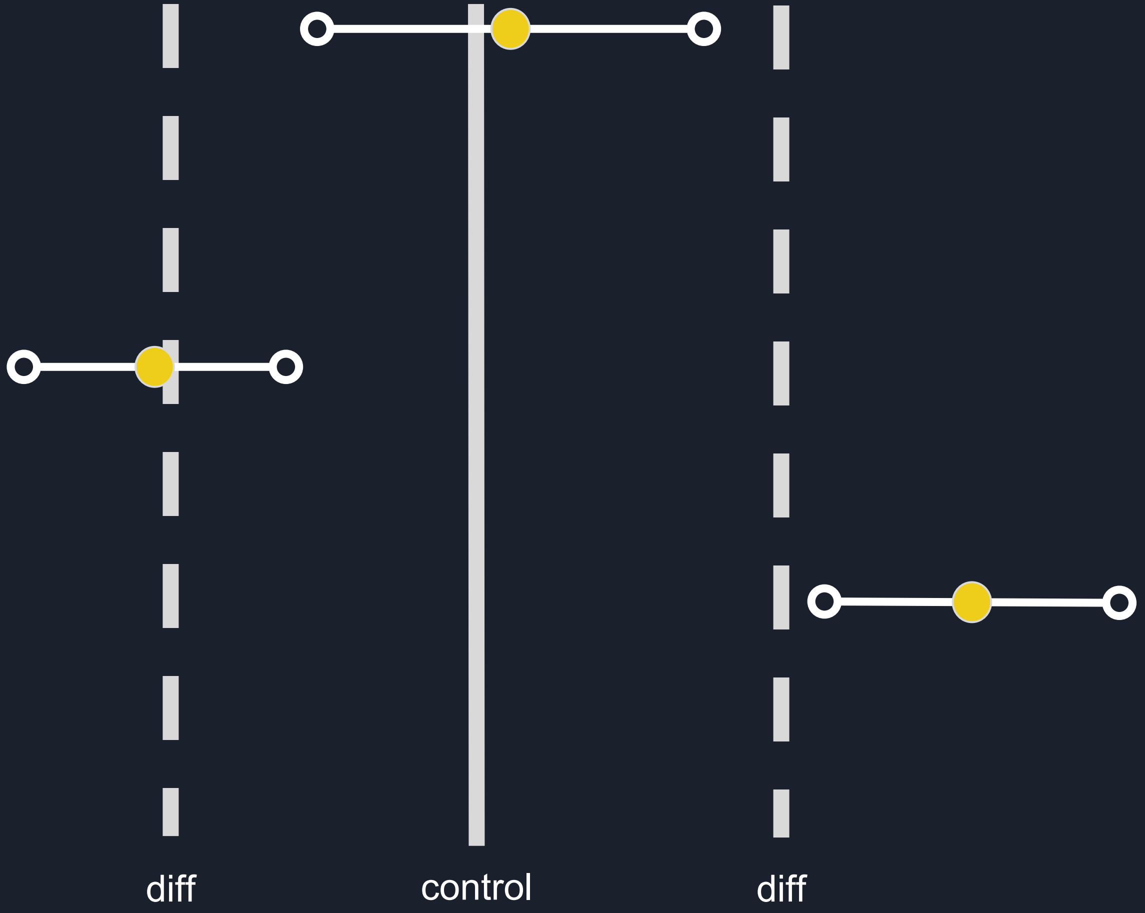


$$n = 16 * 2 * \frac{\sigma^{2}}{\delta^{2}} = 32 * \frac{.499^{2}}{.03^{2}} = 8,853 \textit{ total users}$$

Interpreting Results

Control Probability	49.68% (i.e. 2015/4056)
Treatment Probability	52.96% (i.e. 2345/4428)
Absolute Difference	3.28%
Test Statistic, Z	3.02
Standard Error	0.75%
Confidence Interval	(51.49%, 54.43%)

Should I Launch?



Lessons Learned

Partitioning Users Part II

- Multi-level Bucketing
- Partitioning against multiple treatments

Newness Effects

- Primacy
- Novelty

Metrics Matter

- Generally Applicable
- Score Cards
- Validate Your Metrics
- Have Invariant Metrics

Experiment Often

- Long running experiments have limits
- Filtered and targeted experiments
- Don’t assume it’s trending
- Test everything you can

They’re all nails to me!

- User Surveys
- Research Panels

Questions/Comments/Criticisms

Resources:
Sample Sizing for everyone:
http://www.evanmiller.org/ab-testing/sample-size.html

First Paper:
http://www.exp-platform.com/Documents/GuideControlledExperiments.pdf

Hashing Functions Comparison:
https://github.com/rurban/smhasher