Online Versus Offline Experimentation

Roger Longbotham[1]

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

Abstract

Most statisticians are taught Design of Experiments (DOE) as part of their education. This field was developed before the internet and deals almost exclusively with offline experiments. Having a career with extensive experience in both online and offline experimentation, my objective with this article is to point out how online experimentation is different from offline experimentation for those who have had some training in DOE. Since many statisticians may be considering conducting online experiments, I end with advice on what support one would need to do that.

Key Words: experimentation, online, offline, design of experiments, DOE, AB testing

Introduction

First, an overview of online experimentation. Almost all large online companies run experiments - Google, Amazon, Facebook, LinkedIn, Microsoft, Etsy, Booking.com, Expedia, Yahoo!, etc. These companies run many experiments every year and most sites have multiple experiments running at any one time. These experiments are not in the usability lab but include most or all of their daily customers. The earliest online experiments I am aware of were run by Amazon in the late 90's, although other sites could have been doing experiments earlier. The practice of online experimentation goes by many names: A/B Testing, Multivariate tests, parallel tests, bucket tests, split tests, online field experiments, and more.

What can be tested on a website? Almost any change someone wants to make to the site can, and should, be tested. This could be simple changes like the color or font of some of the text, which image to show, size of images, etc. Here is a more complete list of potential changes a site could test:

Changes that can be tested

User Interface (UI)

- Any change to text or content
- Design of page
- Colors, fonts, images and image size, etc.

[1] Roger Longbotham was instrumental in conducting hundreds of experiments offline while an employee at Rockwell International and as Technical Director for QualPro, Inc. He has also been involved in more than a thousand online experiments at Amazon.com, Microsoft (Bing, MSN, Office online, etc.) and other companies.

User Support Options

- Where to place user support options
- Which user support options (phone number, chat, link to support FAQs, etc.)

Algorithms

- Search algorithms
- Personalization and recommendation algorithms
- Content importance algorithms

Apps

Device specific changes

- Changes to the site that only affect tablets, phones, etc.

Underlying codebase

- Changes to content management system, etc.

Therefore, online experiments can and should include any visual change that a visitor may notice as well as "behind-the-scenes" changes that a visitor would not notice such as search results, recommendations etc. Even bug fixes can be tested if it is uncertain what the full extent of the effects of the bug fixes will be. In online experiments, the website without the change is to be compared to the website with the change. Common terminology for the website without the change is the control or version A. You may consider this the default, or baseline. The website with the change is the treatment or version B. The simplest online experiment is just one treatment versus control, but many experiments will have multiple treatments or versions that are to be compared to control.

Responses from Online Experiments

The data that is collected and metrics that are formed to judge the effect of a treatment vary according to the objectives of the site and the sophistication of the site instrumentation. For example, an online retailer is keenly interested in whether a visitor makes a purchase or not and how large a purchase they make. A content site that primarily has articles (or pictures or videos) for visitors is interested in how many articles the visitor reads, how long they stay on the site and how often they return to the site. All sites should be measuring some basic metrics

Engagement

- Number of visitors
- Activity of each visitor
    - Number of pages viewed
    - Number of sessions[2]

[2] A session is defined as a set of activities/events by a visitor if there is no gap in the events of 30 minutes or more. If there is a 30 minute gap, a new session is started. The length of a session is the last time stamp minus the first time stamp.

  - Average length of session
- Goals achieved, depends on the site. Some common ones:
  - Number of downloads
  - Sign up for newsletter

Loyalty

- Number of days visitor returns per month
- Number of sessions per month

In addition, metrics should be broken down by some easily collected information from the user agent such as country, language, browser, operating system and more. The response metrics an online site collects for an experiment can easily number in the hundreds.

The most common unit of randomization is the visitor, or user.

Major Differences Between Online and Offline Experimentation

Dimension	Online	Offline
Sample Size	• Thousands of experimental units (EUs) to many millions • In most cases EU is a visitor to a website	• Often quite small (less than 20 to dozens to thousands) but some exceptions (marketing experiments) • Fewer EUs available
Signal-to-Noise Ratio	• Low S-N ratio • High noise and want power to see a relatively small (1-2%) change	• Usually higher S-N • Often can get fairly low variability metrics and precise measurements • Commonly looking for larger changes (10-50%)
Standardization and Costs	• Standardization: • Software platforms automate many of the elements of experimentation: – randomization – collection of data – turn on/off – analysis, etc. • EU cost is free or minimal (standardized to be visitor)	• Standardization: • Not automated – design – experimental unit – randomization – data collection – analysis – monitoring (diagnostics) – control during experiment • EU cost is anywhere from free to quite expensive (e.g. in manufacturing)
Nature of Variation	• Large common cause variation – visitor to visitor. • Robots and other outliers. • Very large temporal instability. • Most metrics are highly skewed. (Normally need >1000 EUs for	• Small to large common cause (within) variation. • Few outliers which can often be attributed to special causes and corrected. • Often assumed to be stable environment.

	CLT to hold. Many more needed for some metrics)	• Often metrics are not far from Normal (i.e. need <100 EUs for CLT to hold)
Design of Experiments	• Mantra: keep it simple. Most practitioners test one factor (or idea) at a time. • May have large number of people in a company running experiments and attempts to coordinate/combine tests of multiple factors is seen as more trouble than worth (in most cases). • Must run all treatment combinations concurrently due to temporal instability.	• Experiments are usually much more manual and costly to set up and sometimes costly for experimental units so need to get the most information from each experiment. • The incremental cost from adding more factors to test is small relative to the cost of running the experiment. • Often can run treatments combinations sequentially.
Multi-factor Experiments	• Tests of more than one factor in an experiment is easily carried out as a full factorial with EUs being independently randomized within each factor.	• Often need to run experiments with fewer treatment combinations than a full factorial due to – sample size restrictions, – to get better estimate of within standard deviation, – complexity of running experiment with many TCs is logistically challenging.
Analysis	• Almost always automated. Can be simple statistical t tests or more complex tests such as permutation tests, etc. • May have hundreds of metrics per experiment.	• Ideally simple. Usually ad hoc (i.e. set up just for each experiment.) • Relative few metrics/experiment (usually <10)
What Can Go Wrong	• Robots • Triggering (logic that says who is in the experiment) • Other tests at same time with same EUs (interactions)	• Many sources of potential problems – Test design – Test execution – People not carrying out their recipe – Measurements
Who Runs Experiments	• Many people in organization. Anyone who wants to make a change to the website • No one else needed to carry it out.	• Experiment set up and analysis by expert. • May need help from many people to carry out.
Number of Experiments	• Many • Some websites have many thousands/year. (Some >10,000/yr)	• Relatively few for most organizations.
Focus of Statistical Development	• Increase power.	• Creative designs to

• Able to run more experiments with current EUs without interfering with each other. • Control/minimize FDR while maintaining power.	– Test more factors with fewer EUs, – Get exact information needed (e.g. regarding interactions), etc.

What Do You Need To Run Online Experiments?

If you are comfortable running offline experiments, what else is needed to run online experiments?

- **Need a software platform**. You may have a long development cycle unless you get a start from a vendor. Ad hoc experiments are much more costly than with a platform, so most sites will develop a platform for continuous experimentation.
- **Can't do it alone**. Need support from programmers, computer scientists to set up and maintain experimentation platform. (Some of this could be provided by a vendor, but you'll need in-house expertise as well.)
- **Software skills** to do off-line/ad hoc analysis and investigate special causes.
- **Conduct training**. Large websites may have many (hundreds) of people who occasionally or constantly run experiments so need to conduct training so they are competent to run, monitor and interpret results correctly.

Bibliography

Deng, Alex, Ya Xu, Ron Kohavi, and Toby Walker. 2013. "Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data." WSDM 2013.

Deng, Shaojie, Roger Longbotham, Toby Walker, and Ya Xu. 2011. "Choice of Randomization Unit in Online Controlled Experiment." Joint Statistical Meetings Proceedings. 4866-4877.

Hochberg, Yosef, and Yoav Benjamini. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing Series B." Journal of the Royal Statistical Society 57 (1): 289-300.

Kaushik, Avinash. 2006. "Experimentation and Testing: A Primer." Occam's Razor. May 22. Accessed 2008. http://www.kaushik.net/avinash/2006/05/experimentation-and-testing-a-primer.html. 10

Kohavi, Ron, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. "Trustworthy online controlled experiments: Five puzzling outcomes explained." Proceedings of the 18th Conference on Knowledge Discovery and Data Mining. http://bit.ly/expPuzzling.

Kohavi, Ron, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. "Online Controlled Experiments at Large Scale." KDD 2013: Proceedings of the 19th

ACM SIGKDD international conference on Knowledge discovery and data mining. http://bit.ly/ExPScale.

Kohavi, Ron, Alex Deng, Roger Longbotham, and Ya Xu. 2014. "Seven Rules of Thumb for Web Site." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). http://bit.ly/expRulesOfThumb.

Kohavi, Ron, Roger Longbotham, and Toby Walker. 2010. "Online Experiments: Practical Lessons." IEEE Computer, September: 82-85. http://bit.ly/expPracticalLessons.

Kohavi, Ron, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. "Controlled experiments on the web: survey and practical guide." Data Mining and Knowledge Discovery 18: 140-181. http://bit.ly/expSurvey.

McFarland, Colin. 2012. Experiment!: Website conversion rate optimization with A/B and multivariate. New Riders.

Schrage, Michael. 2014. The Innovator's Hypothesis: How Cheap Experiments Are Worth More than Good Ideas. MIT Press.

Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. "Overlapping Experiment Infrastructure: More, Better, Faster Experimentation." KDD.