Can I Take a Peek?

Continuous Monitoring of A/B tests

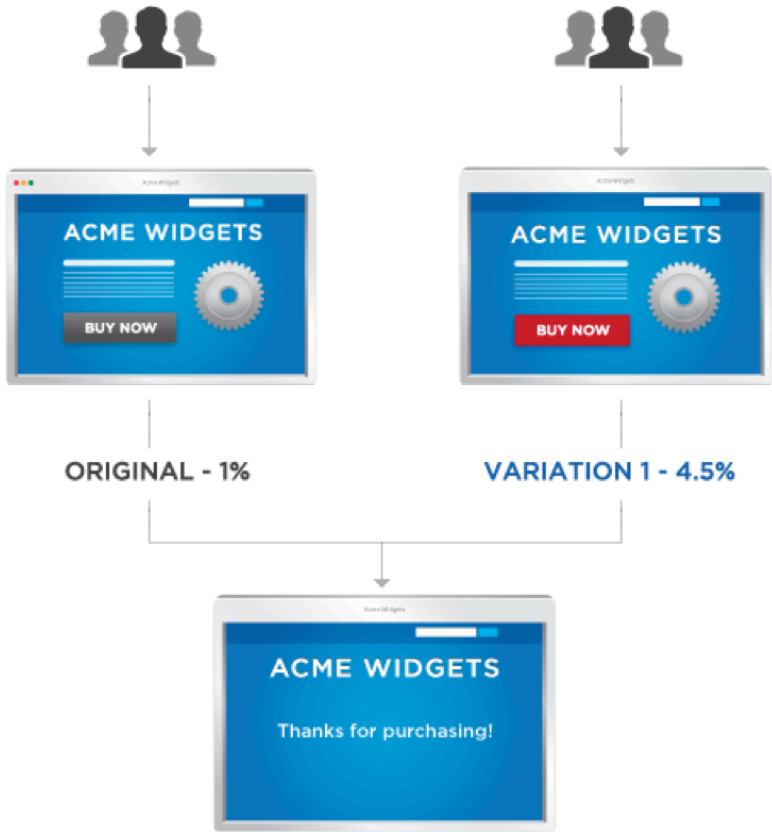
Leo Pekelis | David Walsh | *Ramesh Johari*
Stanford University / Optimizely
ramesh.johari@stanford.edu

18 May 2015

Background: Online A/B Testing

What is A/B testing?

- *A/B testing* = randomized controlled trials used by technology companies and web applications
- Typical use case: comparing versions of a web page
- Question: ***Which yields higher conversion rate?***



Fixed-horizon hypothesis testing

Null hypothesis:
Original conversion rate = new conversion rate

Correct approach:

1. Choose desired false positive probability bound α.
2. Choose minimum detectable effect (MDE).
3. Choose desired false negative probability bound β (at MDE).

$\Longrightarrow$ Yields required *sample size* N

Run experiment and reject null if p-value $p_N \leq \alpha$.

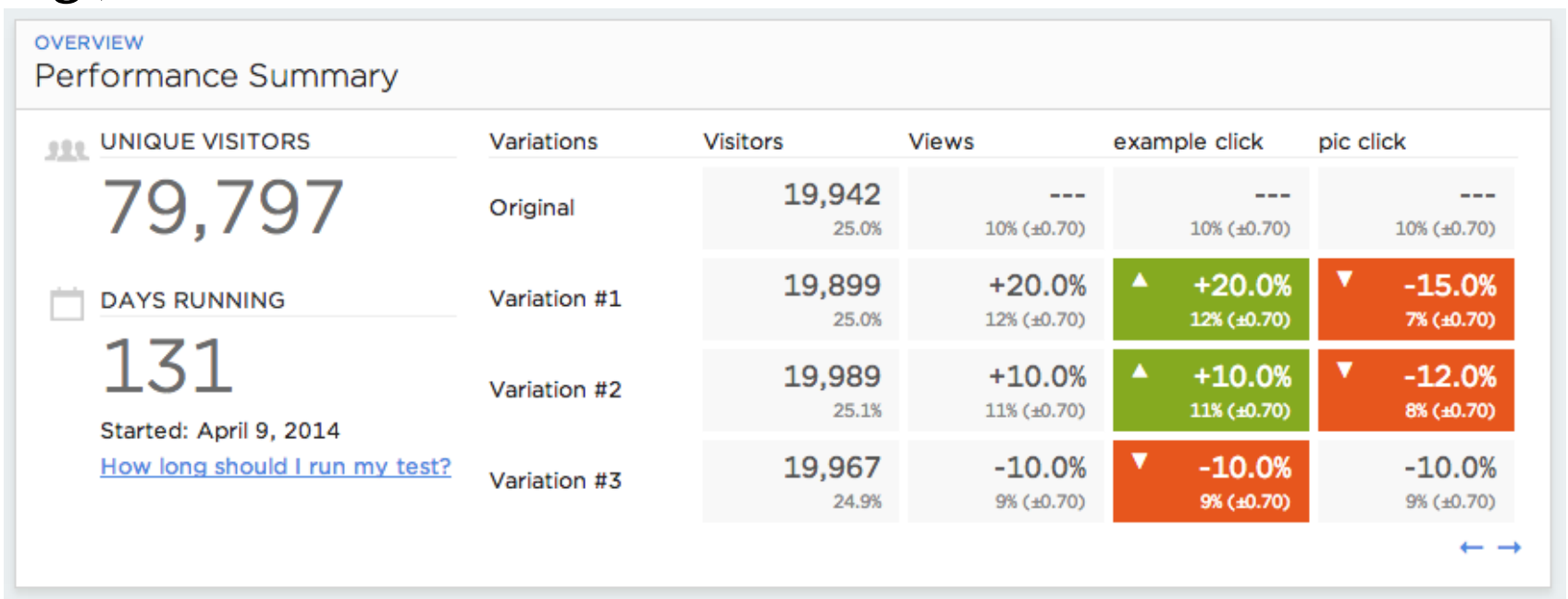
Continuous Monitoring

Continuous monitoring

In practice:

Technology makes it convenient to
continuously monitor tests!

E.g., results matrix:



The problem with peeking

Why is "peeking" at results problematic?

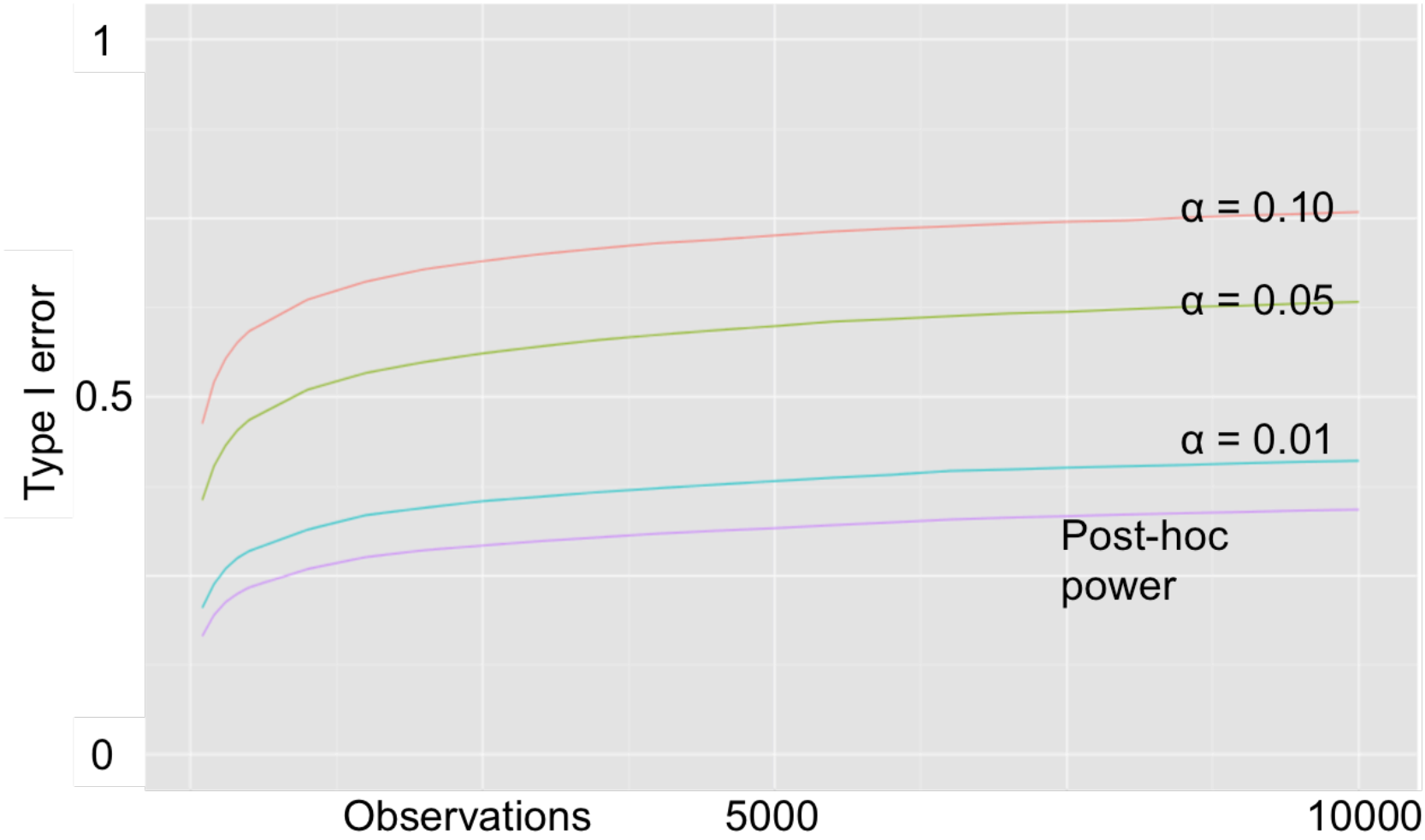
The problem with peeking

Why is "peeking" at results problematic?

Because the decision to reject
may be influenced by the sample path.

The problem with peeking

Even on finite horizons,
naive policies significantly inflate Type I errors:



Always Valid Statistics

Always valid p-values

A desired approach is one which decouples a user's interaction model from providing statistically valid results.

Goal:

- A user should be able to look at their results **whenever** they want.
- The p-value at that time should give valid type I error control.

Always valid p-values

Definition

A **(fixed-horizon) p-value** process is an adapted sequence p_n such that for all n and all $t \in [0, 1]$:

$$\mathbb{P}_0(p_n \leq t) \leq t.$$

Always valid p-values

Definition

A **(fixed-horizon) p-value** process is an adapted sequence p_n such that for all n and all $t \in [0, 1]$:

$$\mathbb{P}_0(p_n \leq t) \leq t.$$

Definition

A p-value process is **always valid** if for any stopping time T and all $t \in [0, 1]$:

$$\mathbb{P}_0(p_T \leq t) \leq t.$$

Always valid p-values

Definition

A **(fixed-horizon) p-value** process is an adapted sequence p_n such that for all n and all $t \in [0, 1]$:

$$\mathbb{P}_0(p_n \leq t) \leq t.$$

Definition

A p-value process is **always valid** if for any stopping time T and all $t \in [0, 1]$:

$$\mathbb{P}_0(p_T \leq t) \leq t.$$

- Allows user to choose T in a data-dependent fashion.

Do always valid p-values exist?

Trivially yes: Let $X \sim U(0, 1)$, and let $p_n = X$ for all n.

Do always valid p-values exist?

Trivially yes: Let $X \sim U(0, 1)$, and let $p_n = X$ for all n.

Why is this undesirable?

Do always valid p-values exist?

Trivially yes: Let $X \sim U(0, 1)$, and let $p_n = X$ for all n.

Why is this undesirable?

Low power:
Not guaranteed to detect a difference if one exists.

Tests of Power One

Sequential tests of power one

Use *sequential tests of power one* to show that "good" always valid p-values exist.

Definition

A **sequential test of power one** is a family of stopping times $\{ T_\alpha : \alpha \in [0,1] \}$ such that:

1. T_α is non-increasing in α;
2. $\mathbb{P}_0(T_\alpha < \infty) \leq \alpha$;
3. $P_\theta(T_\alpha < \infty) = 1$ for any $\theta \neq 0$.

[Analyzed by Robbins and Siegmund in 1970s; further developed by Lai et al.]

Constructing always valid p-values

Theorem

Suppose $\{T_\alpha\}$ *is a sequential test of power one. Define:*

$$p_n = \inf\{\alpha : T_\alpha \leq n\}.$$

Then the resulting process is an always valid p-value process.

Constructing always valid p-values

Theorem

Suppose $\{T_\alpha\}$ is a sequential test of power one. Define:

$$p_n = \inf\{\alpha : T_\alpha \leq n\}.$$

Then the resulting process is an always valid p-value process.

In other words:

p-value = *smallest α such that α-level test would have rejected by now.*

Stopping when $p_n \leq \alpha$ recovers the original sequential test $\implies$ ***power one***

Proof of theorem

Step 1. p_n is decreasing.

Proof of theorem

Step 1. p_n is decreasing.

Step 2. Thus p_∞ exists a.s.

Proof of theorem

Step 1. p_n is decreasing.

Step 2. Thus p_∞ exists a.s.

Step 3. For fixed $\alpha > t$, the event $\{T_\alpha < \infty\}$ contains the event $\{p_\infty \leq t\}$, so:

$$\mathbb{P}_0(p_\infty \leq t) \leq \mathbb{P}_0(T_\alpha < \infty) \leq \alpha.$$

Proof of theorem

Step 1. p_n is decreasing.

Step 2. Thus p_∞ exists a.s.

Step 3. For fixed $\alpha > t$, the event $\{T_\alpha < \infty\}$ contains the event $\{p_\infty \leq t\}$, so:

$$\mathbb{P}_0(p_\infty \leq t) \leq \mathbb{P}_0(T_\alpha < \infty) \leq \alpha.$$

Step 4. Thus taking $\alpha \to t$, for any stopping time T:

$$\mathbb{P}_0(p_T \leq t) \leq \mathbb{P}_0(p_\infty \leq t) \leq t.$$

Example with $N(\mu, 1)$ data

Tests of power one exist, so always valid p-values exist.

E.g. consider $N(\mu, 1)$ data, and testing:

$$H_0 \ : \ \mu = \mu_0$$
$$H_1 \ : \ \mu \neq \mu_0$$

Notation: Let $L_n(\mu, \mu_0; \overline{x}, n)$ be LR of μ against μ_0, given n observations with sample mean $\overline{x}$.

Example with $N(\mu, 1)$ data

Let $G \sim N(0, \sigma^2)$, and consider:

$$\mathbf{L}_n = \int L_n(\mu, \mu_0; \overline{X}_n, n)\, dG(\mu).$$

Define:

$$T_\alpha = \inf\left\{ n : \mathbf{L}_n \geq \frac{1}{\alpha} \right\}.$$

Proposition (Robbins and Siegmund)

$\{T_\alpha\}$ is a sequential test of power one.

[Proof: Optional stopping on $\mathbf{L}_n$.]

Run length

No free lunch?

What do we give up in return for continuous monitoring?

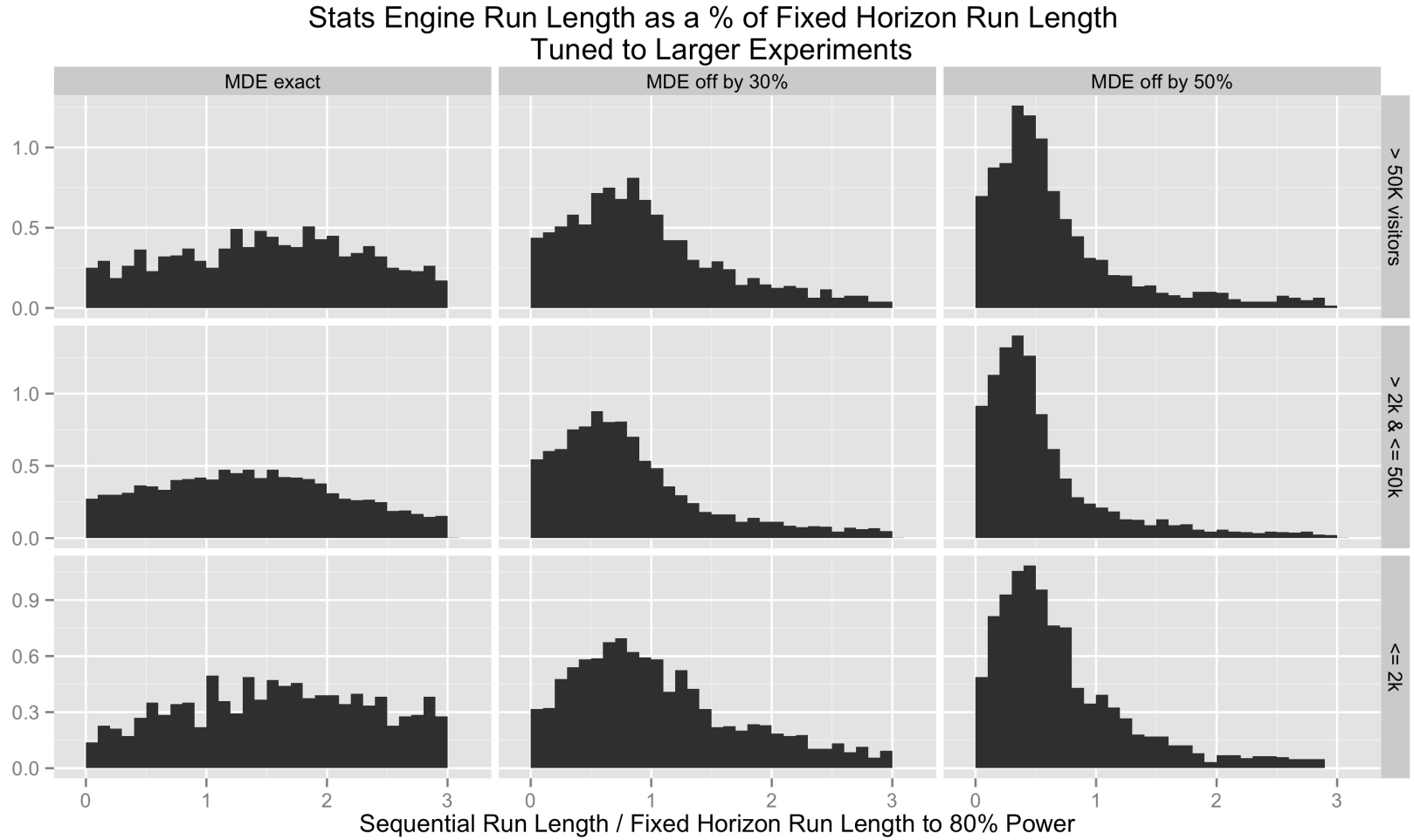
- If the effect size is known in advance,
  should only be better!
- In practice, we don't know the effect size in advance.

The test we designed does not assume knowledge of the effect size.

We compare our test to a fixed horizon test,
using data from Optimizely.

Run lengths on Optimizely

Our results show robustness to not knowing the effect size:



Run lengths: Interpretation

Our results show robustness to not knowing the effect size.

Intuition:

- Detecting an effect of size Δ takes a run length proportional to $1/\Delta^2$
- So the penalty for guessing wrong about δ is very high!
  - An MDE that is 2x too small $\implies$ run length that is 4x too long

Conclusions

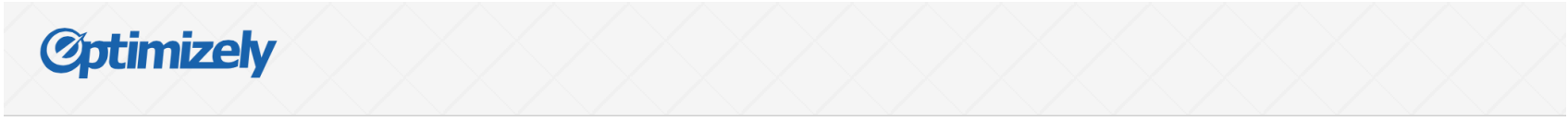
Experimentation in the Internet age

Rapid innovation in
information & communication technology
has **democratized the scientific method**.

Our goal: "adapt" statistical methodology to
fit what the user is trying to do.
Additional results:

- Always valid confidence intervals
- Always valid multiple testing corrections

Optimizely Stats Engine



USING OPTIMIZELY

Statistics for the Internet Age: The Story Behind Optimizely's New Stats Engine

By Leonid Pekelis

- The ideas presented in this talk were released to Optimizely's customers on January 20, 2015
- Provides both always valid p-values and multiple testing corrections