# Choice of Experiment Unit for Controlled Experiments on the Web

Shaojie Deng

One Microsoft Way
Microsoft

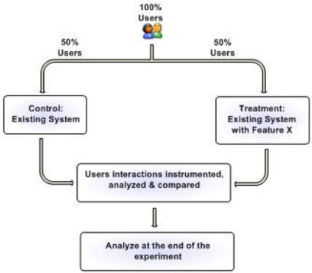
Joint work with Roger Longbotham, Toby Walker and Ya Xu

# Outline

# Agenda

# Introduction

- Controlled experiment or randomized experiment is the best-known scientific method for establishing causality between a feature and its effects.
- The basic methodology is to expose a percentage of users to a new treatment, measure the effect on metrics of interest, and run statistical tests to determine whether the differences are statistically significant, thus establish causality.
- It is easy to collect data on web quickly and at low cost. Web provides an unprecedented opportunity for us to use the power of controlled experiment to test and evaluate ideas quickly.



# Terminology

- **Experiment Unit**: The unit on which the randomization is applied. Also called randomization unit. The most commonly used experiment unit is user or its surrogate such as cookie or user ID. Page view has also been used in practice.

# Terminology

- **Experiment Unit**: The unit on which the randomization is applied. Also called randomization unit. The most commonly used experiment unit is user or its surrogate such as cookie or user ID. Page view has also been used in practice.
- **Metric**: An statistic stands for a concept that the experiment designer wants to evaluate. Common metrics include clicks per user, sessions per user, click through rate, coverage rate, revenue per user etc.

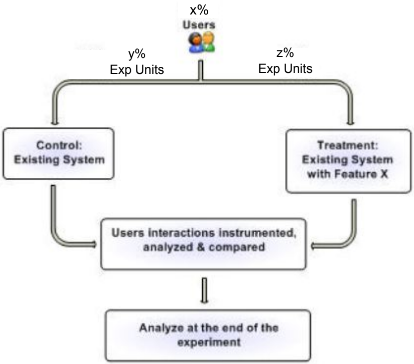
# Terminology

- **Experiment Unit**: The unit on which the randomization is applied. Also called randomization unit. The most commonly used experiment unit is user or its surrogate such as cookie or user ID. Page view has also been used in practice.
- **Metric**: An statistic stands for a concept that the experiment designer wants to evaluate. Common metrics include clicks per user, sessions per user, click through rate, coverage rate, revenue per user etc.
- **Analysis Unit**: A metric is naturally associated with an analysis unit. For example, per user metric such as clicks per user has the analysis unit user. Click through rate and coverage rate use page view as the analysis unit. The analysis unit associated with a metric is also called the *level* of the metric. The most important two types of metrics are user level metrics and page view level metrics.

# Terminology

- **Experiment Unit**: The unit on which the randomization is applied. Also called randomization unit. The most commonly used experiment unit is user or its surrogate such as cookie or user ID. Page view has also been used in practice.
- **Metric**: An statistic stands for a concept that the experiment designer wants to evaluate. Common metrics include clicks per user, sessions per user, click through rate, coverage rate, revenue per user etc.
- **Analysis Unit**: A metric is naturally associated with an analysis unit. For example, per user metric such as clicks per user has the analysis unit user. Click through rate and coverage rate use page view as the analysis unit. The analysis unit associated with a metric is also called the *level* of the metric. The most important two types of metrics are user level metrics and page view level metrics.
- **Measurement**: A measurement is an observation on an analysis unit. For example, the number of clicks of user $i$ on page view $j$ is a page view level measurement. It can be further rolled up (summed up) to user $i$'s total number of clicks, which is a user level measurement.

- Common experiment unit and analysis unit: user, page view
- Analysis unit always finer than experiment unit.
- Interested in the following combination:
  1. E:User,A:User (Vanilla)
  2. E:User, A: Page View (Delta Method)
  3. E:Page View, A: Page View (This Talk)



# Two Sample T-test

- Given a metric $S$, we can calculate the value of this metric for both control group and treatment group, denoted by $S_c$ and $S_t$. We want to test whether $S_t - S_c = 0$.
- Test statistics $\frac{S_t - S_c}{Var(S_t - S_c)} = \frac{S_t - S_c}{Var(S_t - S_c)}$.
- Central limit theorem guarantees that the test statistic is asymptotically normal. The crux of the problem is to estimate the variance of $S_t - S_c$.

# Agenda

# Vanilla case: E: user+ A: user

- Assume user i.i.d. sampled. Hence user level measurements i.i.d.
- $S_t$ and $S_c$ are sample means of user level measurements.
- Estimate variance of $S_t$ and $S_c$ by sample variance.
- Since user i.i.d. and treatment and control have disjoint set of users, $Var(S_t - S_c) = Var(S_t) + Var(S_c)$
- Most widely used. Simple and easy to understand.

# E: user+ A: page view

- Common mistake is to treatment page view level measurements as i.i.d.
- Not true, because randomization was on user level, and there is strong between user variance.
- $Var(S_t)$ and $Var(S_c)$ can be estimated via delta method.
- Since user i.i.d. and treatment and control have disjoint set of users, $Var(S_t - S_c) = Var(S_t) + Var(S_c)$.
- Widely used. But not always with delta method correctly applied.

# Delta Method

- Let $X_{i,j}$ be page level measurement for user-$i$'s $j^{th}$ page view, and $K_i$ be user-$i$'s number of page views. User level measurement $(\sum_{i=1}^{K_i} X_{i,j}, K_i), i = 1, \ldots, n$ are i.i.d. Page view level metric is $\overline{X} = \frac{\sum_i \sum_j X_{i,j}}{\sum_i K_i}$.
- By letting $Y_i = \sum_{i=1}^{K_i} X_{i,j}$ and express $\overline{X}$ as $\sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} K_i$, it is then a straightforward application of the delta method to get an asymptotically consistent estimator for $Var\overline{X}$:

$$\frac{1}{n}\left\{\frac{1}{\widehat{\mathbb{E}K_i}^2}\widehat{VarY_i} + \frac{\widehat{\mathbb{E}Y_i}^2}{\widehat{\mathbb{E}K_i}^4}\widehat{VarK_i} - 2\frac{\widehat{\mathbb{E}Y_i}}{\widehat{\mathbb{E}K_i}^3}\widehat{Cov(Y_i,K_i)}\right\}$$

where these "hatted" quantities are the sample mean, variance and covariance.

# Agenda

- Cons: User level metrics not available; variance estimation not straightforward. User experience not consistent, limiting its adoption.
- Pros: Variance reduction for page view level metrics. Better statistical power.
- Page view randomization experiments are good for testing the effect of a treatment on page level metrics where consistent experience is not necessary. For example, test performance metric such as page loading time, test certain ranking algorithm of a search engine, test conversion rate of a promotion page, etc...

# Variance of $S_t - S_c$?

- Remember that we only allocate $x\%$ users for page view level randomization experiments.
- Page views of these $x\%$ users are then split into treatment and control experience.
- $Var(S_t)$ and $Var(S_C)$ can be estimated separately by delta method. But $S_t$ and $S_c$ are not independent and it is unclear how to compute $Var(S_t - S_c)$

# A Bottom-Up Probability Model

- Let $X_{i,j}^{(r)}$ be the page level measurement (e.g. number of clicks on the page) on user $i$'s $j^{th}$ page view in group $r (r = 1, 2$ for control and treatment).
- $X_{i,j}^{(r)}$ has mean $\mu_i$ and variance $\sigma_i^2$ where $(\mu_i, \sigma_i^2)$ can differ from user to user but is fixed for each user. We call this the user effect. Under null hypothesis, control and treatment are the same.
- $K_i$ the total number of page views from user $i$ and $N = \sum_{i=1}^{n} K_i$ be the total number of page views.
- Assume $K_i, i = 1, \ldots, n$ are i.i.d. and independent of $(\mu_i, \sigma_i^2), i = 1, \ldots, n$.(Not always true, need to check this assumption case by case.)
- Each $K_i$ are split into control and treatment, with $K_i^{(1)}$ and $K_i^{(2)}$ for each group. Let $N_{(r)}, r = 1, 2$ be total number of page views.

# Road Map

- Let $\widehat{\sigma^2_{nr}} = n \times \frac{1}{N_r^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{K_i^{(r)}}(X_{i,j}^{(r)} - \overline{X})^2\right)$,
$\widehat{\sigma^2_{dr}} = \left\{\frac{1}{\widehat{\mathbb{E}K_i^{(r)}}^2}\widehat{VarY_i^{(r)}} + \frac{\widehat{\mathbb{E}Y_i^{(r)}}^2}{\widehat{\mathbb{E}K_i^{(r)}}^4}\widehat{VarK_i^{(r)}} - 2\frac{\widehat{\mathbb{E}Y_i^{(r)}}}{\widehat{\mathbb{E}K_i^{(r)}}^3}\widehat{Cov(Y_i^{(r)}, K_i^{(r)})}\right\}$
- We give a consistent estimator of $Var(\overline{X}_1 - \overline{X}_2)$.
- We verify the formula with simulation.

## Theorem

*Let* $w_i^{(r)} = K_i^{(r)} / \sum_{i=1}^n K_i^{(r)}$. *Let* $C_r = \mathbb{E}[(K_i^{(r)})^2]/(\mathbb{E}K_i^{(r)})^2, r = 1, 2,$
$C_x = \mathbb{E}(K_i^{(1)} K_i^{(2)})/(\mathbb{E}K_i^{(1)} \mathbb{E}K_i^{(2)})$. *As* $n \to \infty$.

$$\widehat{\sigma^2_{nr}} \to \frac{1}{\mathbb{E}(K_i^{(r)})}(Var(\mu_i) + \mathbb{E}(\sigma_i^2)).$$

$$\widehat{\sigma^2_{dr}} \to C_r Var(\mu_i) + \mathbb{E}(\sigma_i^2)/\mathbb{E}(K_i^{(r)}).$$

$$nVar(\overline{X}_1 - \overline{X}_2) \to (C_1 + C_2 - 2C_x)Var(\mu_i) + (\frac{1}{\mathbb{E}(K_i^{(1)})} + \frac{1}{\mathbb{E}(K_i^{(2)})})\mathbb{E}(\sigma_i^2).$$

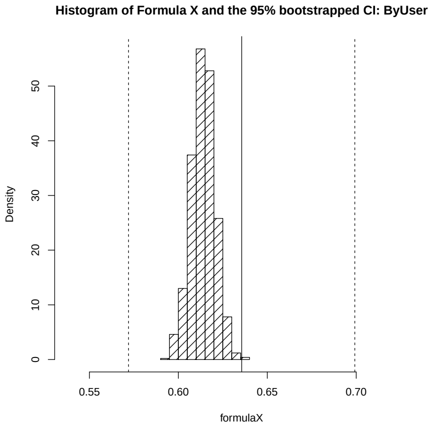
*Moreover,* $C_1 + C_2 - 2C_x = \frac{1}{\mathbb{E}K_i^{(1)}} + \frac{1}{\mathbb{E}K_i^{(2)}}$ . *Therefore*

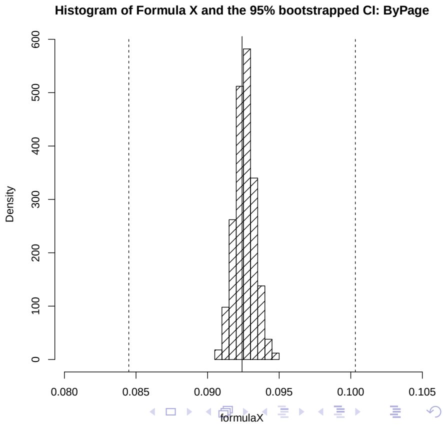
$$nVar(\overline{X}_1 - \overline{X}_2) \to \left(\frac{1}{\mathbb{E}K_i^{(1)}} + \frac{1}{\mathbb{E}K_i^{(2)}}\right)\left(Var\mu_i + \mathbb{E}\sigma_i^2\right).$$

*i.e.,* $\widehat{\sigma^2_{n1}} + \widehat{\sigma^2_{n2}} \to nVar(\overline{X}_1 - \overline{X}_2)$ *(Formula X).*

# Simulation Results: Page click rate

N=1,000,000 users. User page click rate $\mu_i$ from $Beta(0.1, 0.5)$. Sessions of each user from Poisson(2), then Page view for each session Poisson(3). 100,000 simulation runs. 95% CI for true variance from 100 bootstraps.



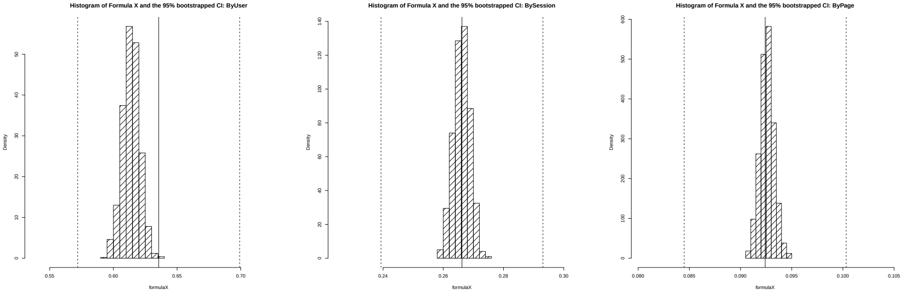
# Agenda

- We presented a solution for statistical testing in the case of page view as both experiment and analysis unit.
- Simulation result shows the performance of the formula X for variance estimation.
- Simulation result shows for the same page view level metrics, using page view as experiment unit leads to much smaller variance.
- We can use other units such as user session (visit) as experiment unit. Using session as both advantage of variance reduction and yet provide reasonable user experience consistency. Formula X can be easily extended.

# Agenda

- $nVar(\overline{X}_1 - \overline{X}_2) \to (C_1 + C_2 - 2C_x)Var(\mu_i) + (\frac{1}{\mathbb{E}(K_i^{(1)}} + \frac{1}{\mathbb{E}(K_i^{(2)}})\mathbb{E}(\sigma_i^2)$ works as long as the assumption $K_i$ independent of $(\mu_i, \sigma_i)$ is valid. It does not depends on the choice of experiment unit.
- $C_1 + C_2 - 2C_x = \frac{1}{\mathbb{E}K_i^{(1)}} + \frac{1}{\mathbb{E}K_i^{(2)}}$ is only true when experiment unit is page view.
- Using $\widehat{\sigma^2_{n1}}$ and $\widehat{\sigma^2_{d1}}$, we can solve $Var(\mu_i)$ and $\mathbb{E}(\sigma_i^2)$. We can then plug them into $(C_1 + C_2 - 2C_x)Var(\mu_i) + (\frac{1}{\mathbb{E}(K_i^{(1)}} + \frac{1}{\mathbb{E}(K_i^{(2)}})\mathbb{E}(\sigma_i^2)$ to get a consistent estimator for $Var(\overline{X}_1 - \overline{X}_2)$.
- In particular, if randomize by user, $C_x = 0$ since one of $K_i^{(1)}$ and $K_i^{(2)}$ will be 0. Above formula reduce to sum of two variance estimations via delta method.
- In general, $C_x > 0$ when experiment unit is finer than user, the finer the experiment unit, the smaller $C_1 + C_2 - C_x$ will be.

# Simulation


Histogram of Formula X and the 95% bootstrapped CI: ByUser
Density
formulaX
0.55
0.60
0.65
0.70
Histogram of Formula X and the 95% bootstrapped CI: BySession
Density
formulaX
0.24
0.26
0.28
0.30
Histogram of Formula X and the 95% bootstrapped CI: ByPage
Density
formulaX
0.080
0.085
0.090
0.095
0.100
0.105

# Fair Pair Experiment

- Reid's team did a randomization by query+user experiment. Control and Treatment ratio 3:1. Total 570,880 users with 3,448,247 page views.
- Treatment shifts the answer insertion by 1 position.
- Variance of win rate from 1,000 bootstrap: 0.0541.
- Variance of win rate from Formula X: 0.0569.
- If use user as experiment unit, variance would be 0.1124.