Trustworthy Analysis of Online A/B Tests

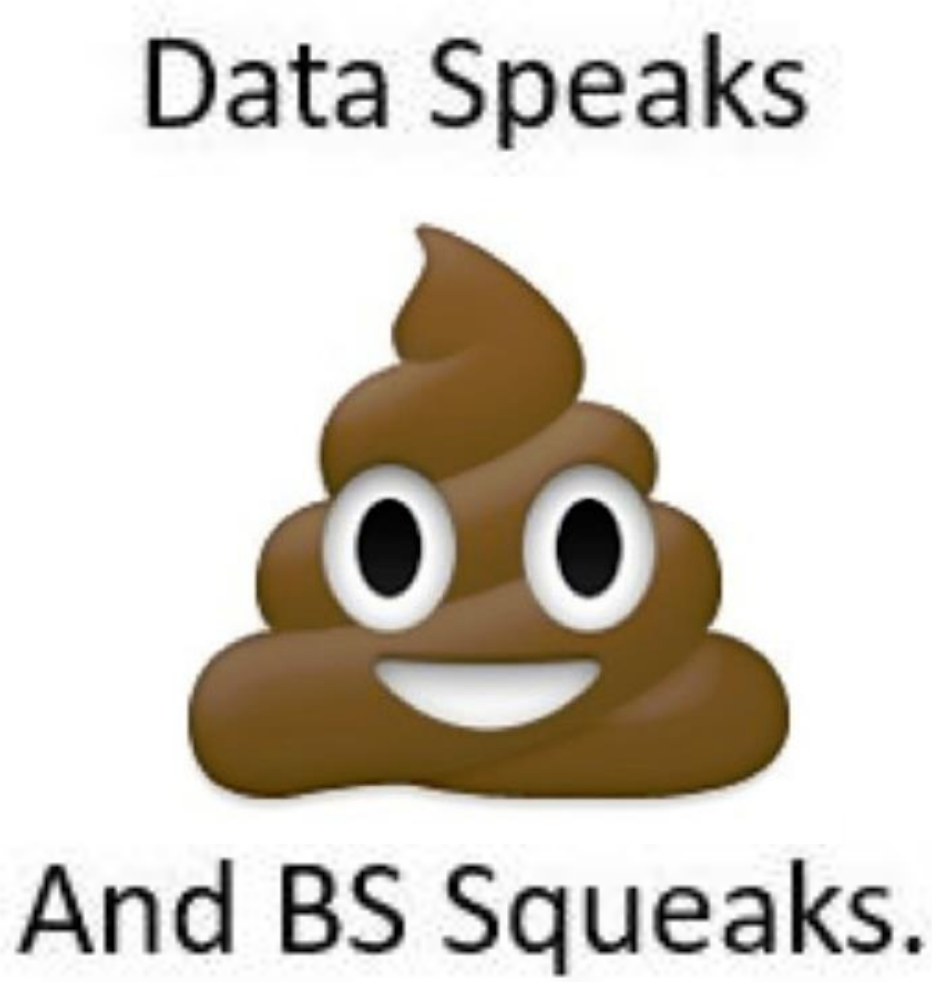
Alex Deng @ Microsoft

Atlantic Causal Inference Conference 2017

About Myself & ExP@Microsoft

- 7 years in ExP team. Statistician by training, nowadays better known as Data Scientists
- Interested in statistical problems combined with engineering challenges
- ExP was founded in 2006 as an incubation team. Now it is the online experimentation platform serves large divisions including Search, Ads, Office, Xbox, Skype and Windows
- Bing alone runs hundreds of experiments concurrently, thousands per month
- Analysis of each experiments need to process several dozen TB of data

Trustworthy analyses: No Bad Science



Today

1. Independence and Randomization
2. Obscured Randomization
3. Empirical Bayes

Independence

Analysis unit (Revenue per ***User,*** Revenue per ***Request***) and Randomization unit

- What justifies i.i.d. assumption?
- Are users i.i.d.?
- Are families i.i.d.?
- Are locations i.i.d.?
- Are organizations i.i.d.?

A short quiz (1min)

There is a large urn full of balls with numbers between 1 to 10

Each time pick one ball, observe the number and then put it back to the urn

Observe a series of numbers. Are these observations independent?

Hint: Independence means knowing previous observations won’t help you predict the next observation

A short quiz (1 min)

Raise your LEFT hand if your answer is YES they are independent

Raise your RIGHT hand if your answer is NO they are not

Answer: Both are correct

Not Independent: Knowing previous numbers help us understand the distribution of numbers in the urn, thus help better predicting the next number

(Bayesian, exchangeable and conditionally independent)

Independent: If we assume the distribution information is public information, e.g. uniform between 1 to 10, then observations are i.i.d. from this distribution

(Frequentist, fixed but unknown information)

Answer: Both are correct

Not Independent: Knowing previous numbers help us understand the distribution of numbers in the urn, thus help better predicting the next number

Independent: If we assume the distribution information is public information, e.g. uniform between 1 to 10, then observations are i.i.d. from this distribution

Independence is not justified by theory, but by *choice!*

Independence and External Validity

Users(or any randomization units) always share some common “*environment*”

If we see this common environment as fixed, then we can assume users i.i.d.

If we expect this environment to also be changing, then not

External Validity: whether result can be generalized outside of the context of the experiment

Randomization Unit Principle

RUP: Randomization unit can typically be treated as independent

- Search Ads experiment randomize on page-views -> pageviews i.i.d.
- Xbox game randomize on game-session -> game-sessions i.i.d.
- Skype randomize on call -> calls i.i.d.

Of course randomization unit has to be chosen appropriate to avoid jarring experience switching

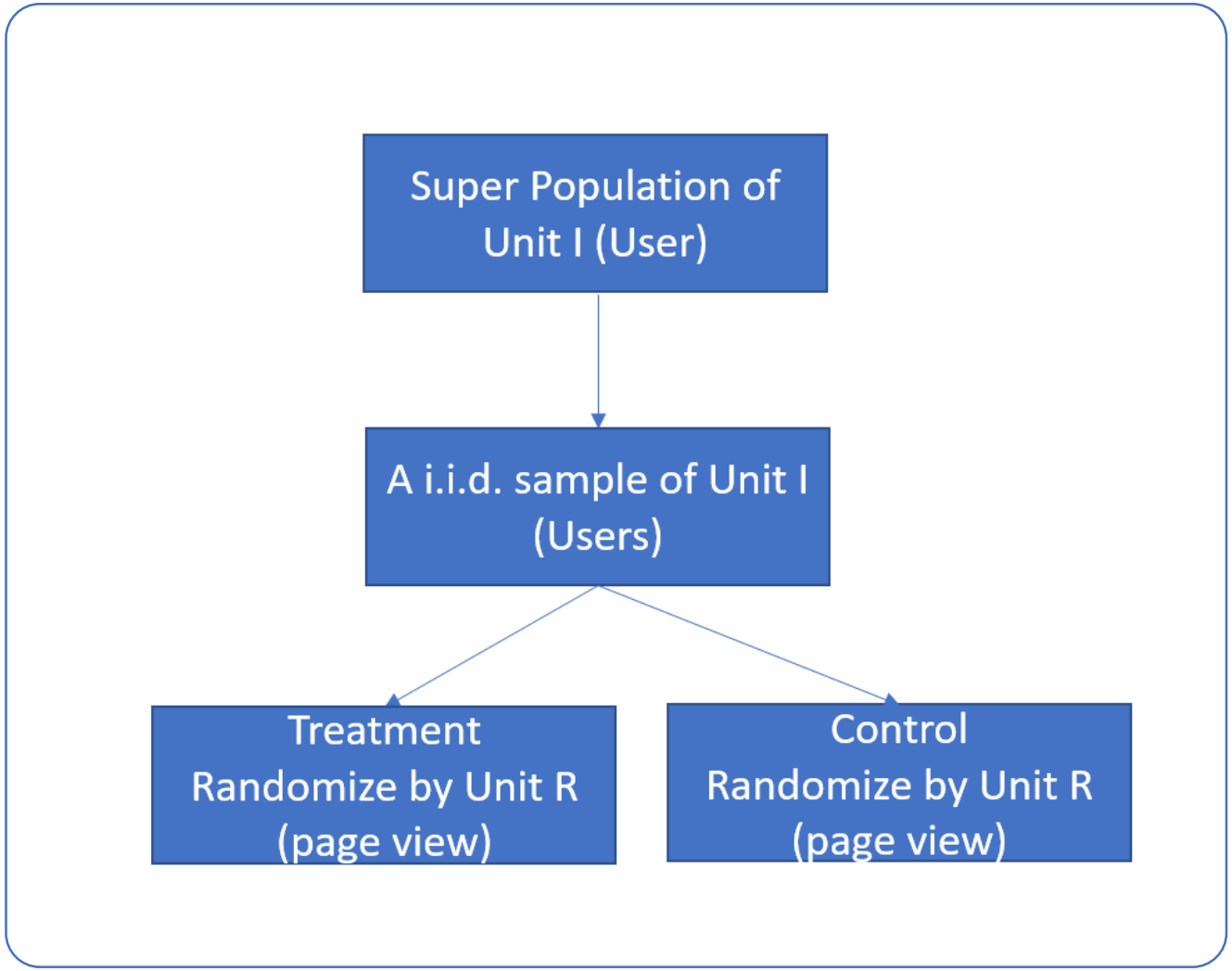
It suggests randomization somehow justifies independence. (seems wrong)

- Assume an independent unit I
- The “traffic” for one experiment is a sample from a super population
- Randomize by randomization unit R (R is a sub-unit of I)

RUP suggests independence justified by randomization. Seems wrong. Is RUP a BS?

RUP has gain faith in A/B testing field for a long time, and it passed all type-I error validation through A/A test (treatment is the same as control, so H_0 is true)

Proof?



- For PATE, we look at $\Delta := \overline{Y_T} - \overline{Y_C}$ (average over unit R, not I)
- Unit i has effect $\tau_i \sim \tau$
- N is sample size of unit I

$$N \times (Var_{RUP}(\Delta) - Var_{True}(\Delta)) \rightarrow O(Var(\tau))$$

- When there is no effect, RUP holds
- When variance of effect is small relative to variance of the metric $\bar{Y}$, RUP provides a good approximation

Complex Randomization

01

Client Experimentation

Mobile app and desktop app need to be working with or without network connection

Randomized assignments are pushed to client every hour

Clients only receive new assignment when connected

Clients apply changes at the next refresh window, e.g. app open or wake from background

02

Social Sharing

Experiment a new way of sharing. Randomized by sharers and treatment got new sharing message/style

Interested in conversion rate

Can share with multiple people

From a receiver perspective, you receive both treatment and control sharing messages depends on who share with you

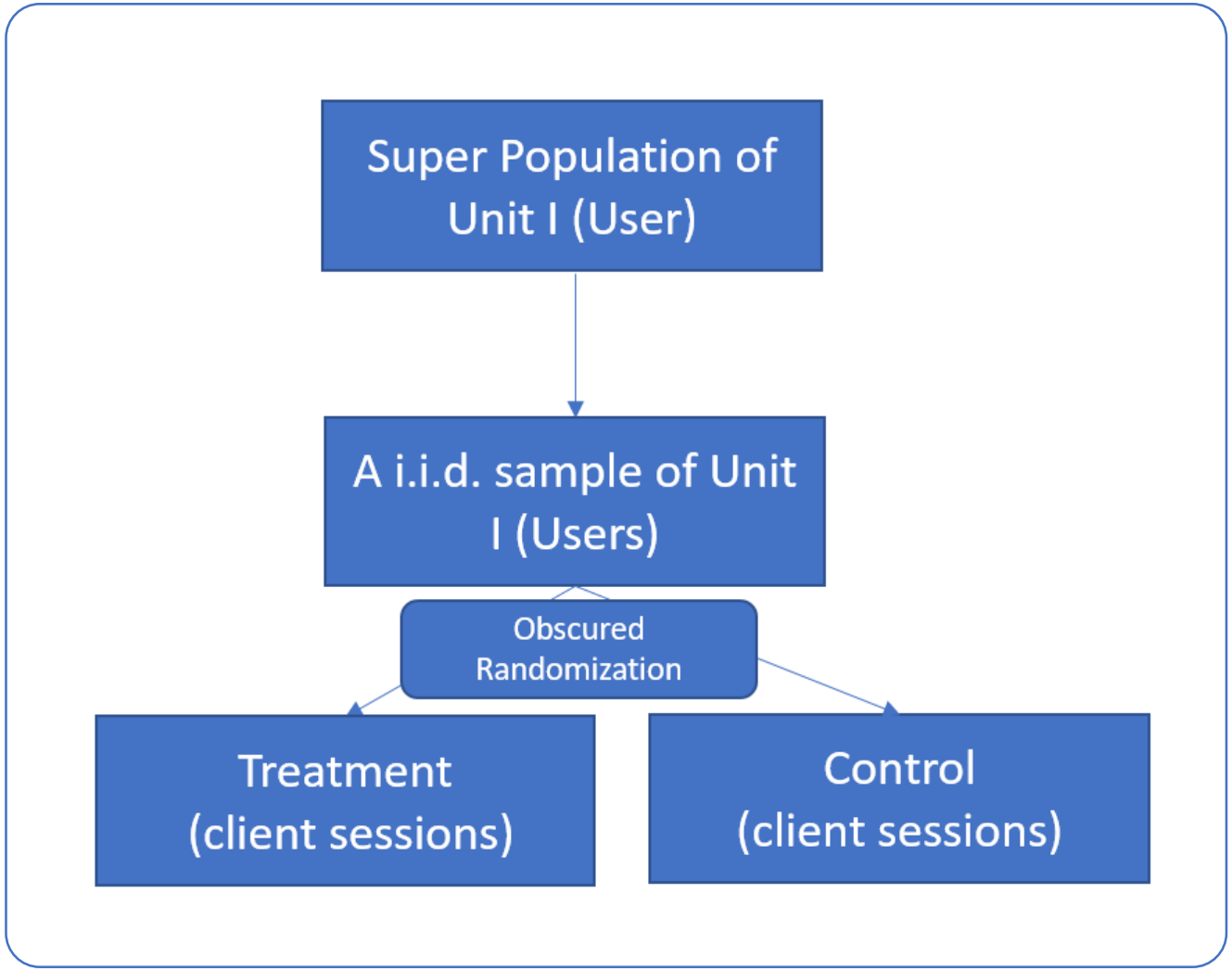
Key observation:

Each independent unit (User) further split into analysis units (sessions)

1. Perfect Randomization: Binomial distribution ($N_T + N_C = N$)
2. One time randomization: $N_T = N$ or $N_C = N$
3. Obscured randomization:
   - We don't know the mechanism
   - Mechanism could vary from user to user! (network availability, behavior, etc.)

Variance of Δ can be derived in a simple form (and computationally scalable)

Technically, it is an interesting application of the ***delta method***



Common Pitfalls

- Continuous monitoring of p-value
- Take p-value as the probability that the null hypothesis is true
- Fail to adjust for multiple comparison/testing

Trustworthy Interpretation

Misconception of P-value

1. P-value is the probability of the null hypothesis being true
2. Studies with the same p-value provide same evidence against null
3. P-value 0.05 means that if you reject the null hypothesis, the probability of a false positive is only 5%

More from Steven Goodman's *"A Dirty Dozen: Twelve P-Value Misconceptions"*

This is NOT a problem of ***experimenter***,
but a problem of the ***platform***

Bayesian Hypothesis Testing

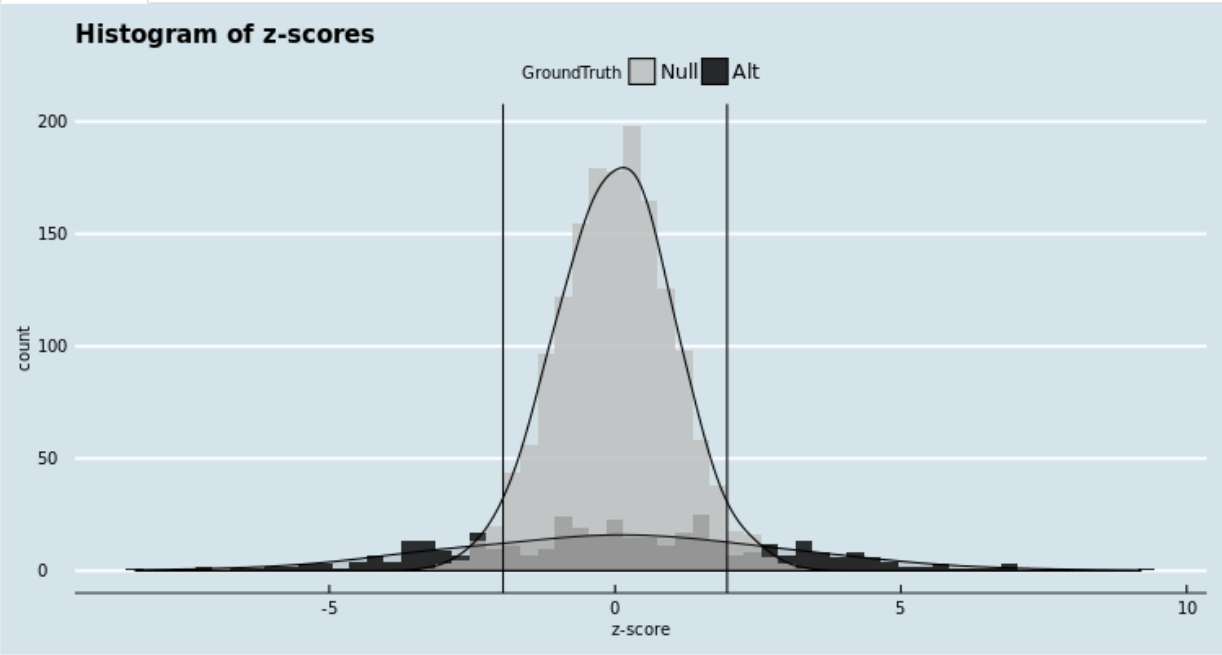
- With P(Null) and P(Alternative), we can truly compute

$$P(\text{Alternative}|\text{Data})$$

- Note P_Null: = 1 – P(Alternative|Data) = P(Null|Data) is similar to False Discover Rate (FDR)
- Allows continuous monitoring: can stop the experiments once FDR is below a threshold
- Also adjusts for (most, but all kinds of) multiple testing. We can compute FDR for a scorecard of thousands of metrics and make decision

Challenge: need two pieces of prior information:

1. P(Null)
2. P(Data|Alternative) depends on the model of “alternative”



We run **thousands** of experiment each year for a scaled partner

- Standard two group model
- P(Null) = p and effect size under alternative follows exponential family/normal
- Estimate parameters using historical data – Empirical Bayes
- Full Bayes/Hierarchical Bayes

Using Rich Historical Experiment Data

Further Challenges

- Are past experiments a good source to learn for new experiments?
- Ignored other rich information:
  - Type of experiment
  - Type of treatment/ideas
  - Team's confidence in the treatment
- Alternative distribution for what?
  - Effect size
  - Z-score/p-value -> local FDR
  - Very different implications, e.g. same p-value, different sample size, should P(Null|Data) be the same?
- Ongoing work
  - Add side information into Empirical Bayes model
  - Additional model assumptions like FDR-regression:

    P(Null|X) ~ logistic(X)

Question?

More at http://exp-platform.com

Slides available at aka.ms/acic07-exp