

# Review: Improve Sensitivity (Power)

## 1. Reduce Variance

- Transform Metrics (dummies, log, capping)
- Paired Design (interleaving, test algorithms)

## 2. Increase Sample Size

- More granular randomization units
- Pooled Control Group (Increase No & Large Control Group)
- Split the sample for the optimal power

## 3. Increase Effect Size ( $\delta$ ) (OECs)

- Trigger Experiments

# Review: Trigger Experiments

- If the experiment only impacts some users, filter out the noises: Units not impacted by the treatments.
- Trigger experiments (random assignment) when certain conditions are satisfied.
- Treatment Effects in the Triggered Population = Treatment Effects on Triggered Population (subset of the population)
- Estimate the Overall Treatment Effects
  - If you improve the revenue by 3% in the triggered population (10% of the population),
  - Overall Treatment Effect = 0 - 3%
    - Depends on how much the rest of traffic contribute to OEC

# Experimentation on a Specific User Segment

- If you improve OEC on a specific user segment
  - A. The lift could be the same for the overall population.
  - B. Even if the treatment effects are large, the overall treatment effects can be very small.
    - This small change may not matter much for a start-up but can still benefit a mature product.
- The experience gained from the triggered experiment on a specific user segment may be generalizable to other features.
  - The algorithms that recommend restaurants may apply to the recommendations of others.

# Always a desire to improve the power

- e.g., If we want to detect a 1% (lift) increase in click-through rate for ads with baseline rate = 1%, how large will the sample be?
  - $n_0 = n_1 = 16 * 0.99 \%^2 / (1\% * 1\%)^2 = 158400$
  - Sample Size (two variants) = 0.32 million
- Tech firms run thousands of experiments per year (Bing, Netflix, Facebook, LinkedIn, Google, etc).
  - to increase the pace of innovation by scaling up the number of experiments.
- Guarantee at least 80% power to detect the difference.
  - To not miss product changes that can have a substantial impact on user experience and revenue.

# S6 Improve Sensitivity II

Reduce Variance: Stratification, Control Variables (CUPED)

Shan Huang, HKU

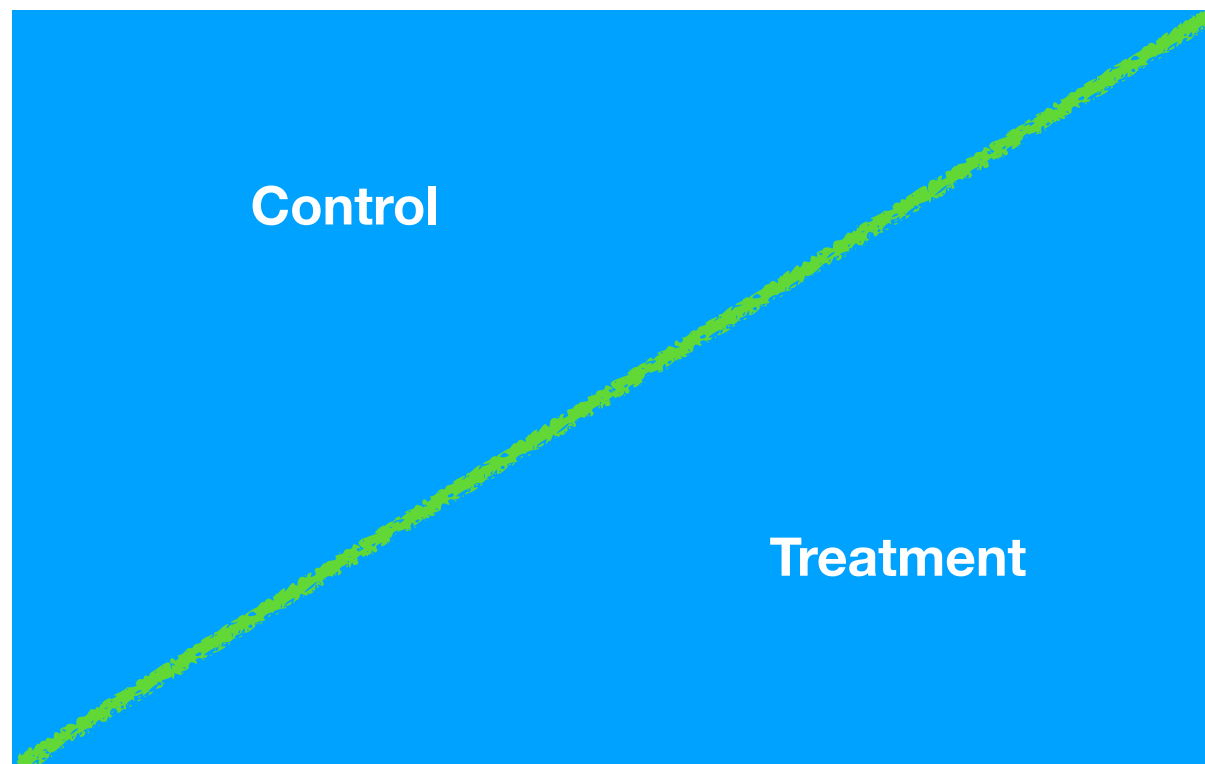
# Reduce Variance $\sigma^2$

- Stratification *at assignment*
  - Post Stratification
- Control Variates
  - Regression with Control Variates
  - CUPED

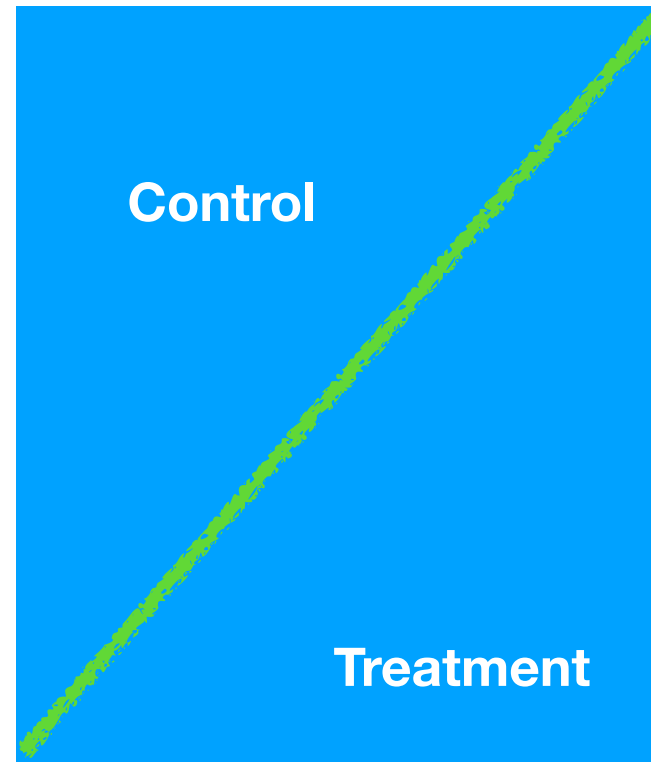
# Stratification

- The most well-known **at-assignment** variance reduction technique.
  - Block Design
- Basic procedure:
  - Divide the **population** into strata
  - Sample from each stratum independently
  - Randomize within each stratum
  - Combine treatment effects from strata to give an overall estimate

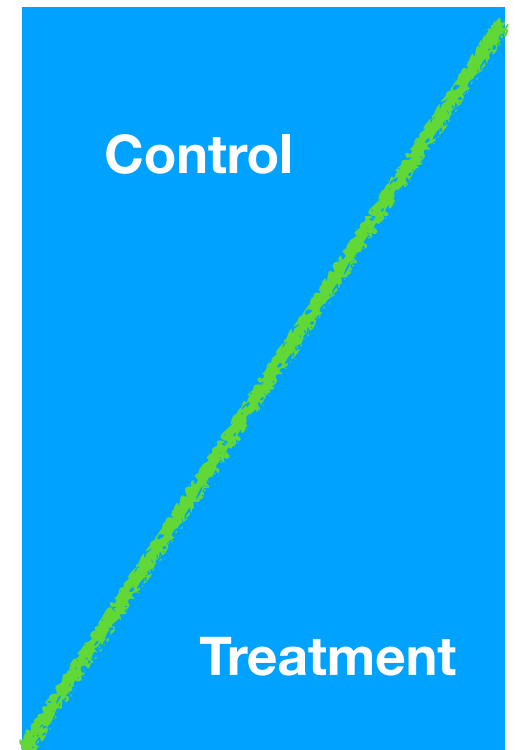
# Block Design



**Stratum 1**



**Stratum 2**

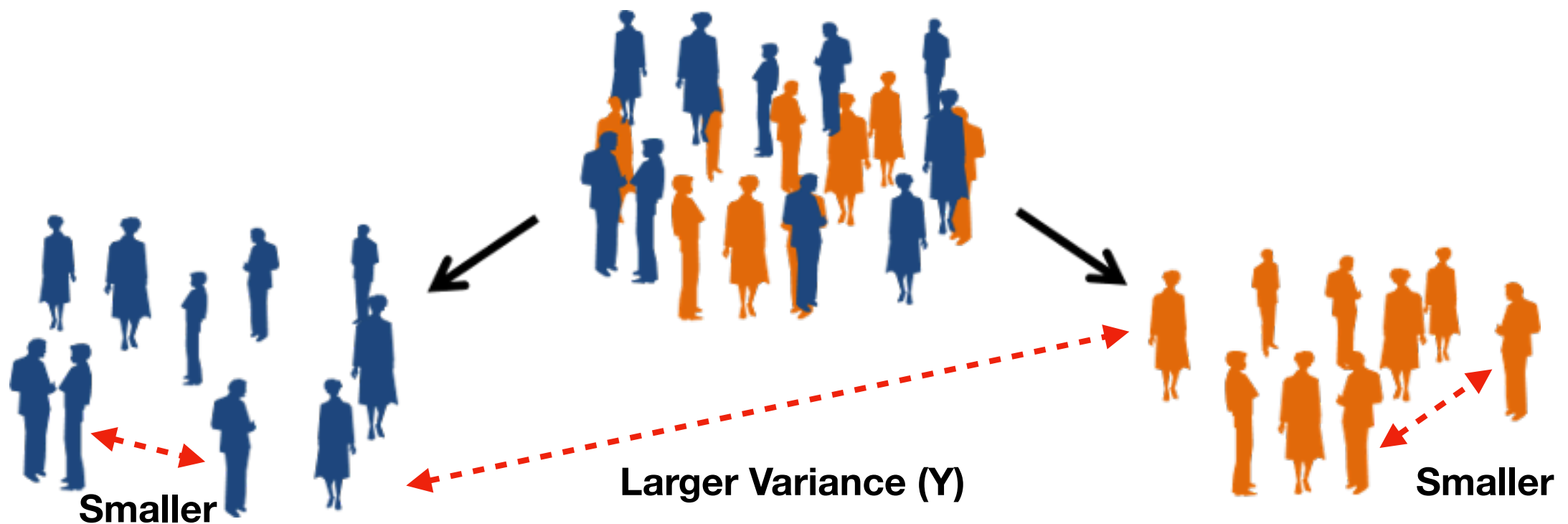


**Stratum 3**



# Stratification

- Idea: Remove between-strata variance in metrics (OECs)
- Choose a categorical variable to divide population into different strata (categories) with different OEC (Y)
- The value of OECs are different across strata.



# How to Divide Population into Strata

- Stratified sampling aims to remove the between-strata variance in OEC (Y).

$$Var(\bar{Y}^s) = Var(\bar{Y}) - \frac{1}{n} \sum_1^K p_k (\mu_k - \mu)^2$$

- Which variable will you choose to divide the population into strata?

OEC is d(purchase) of Taobao

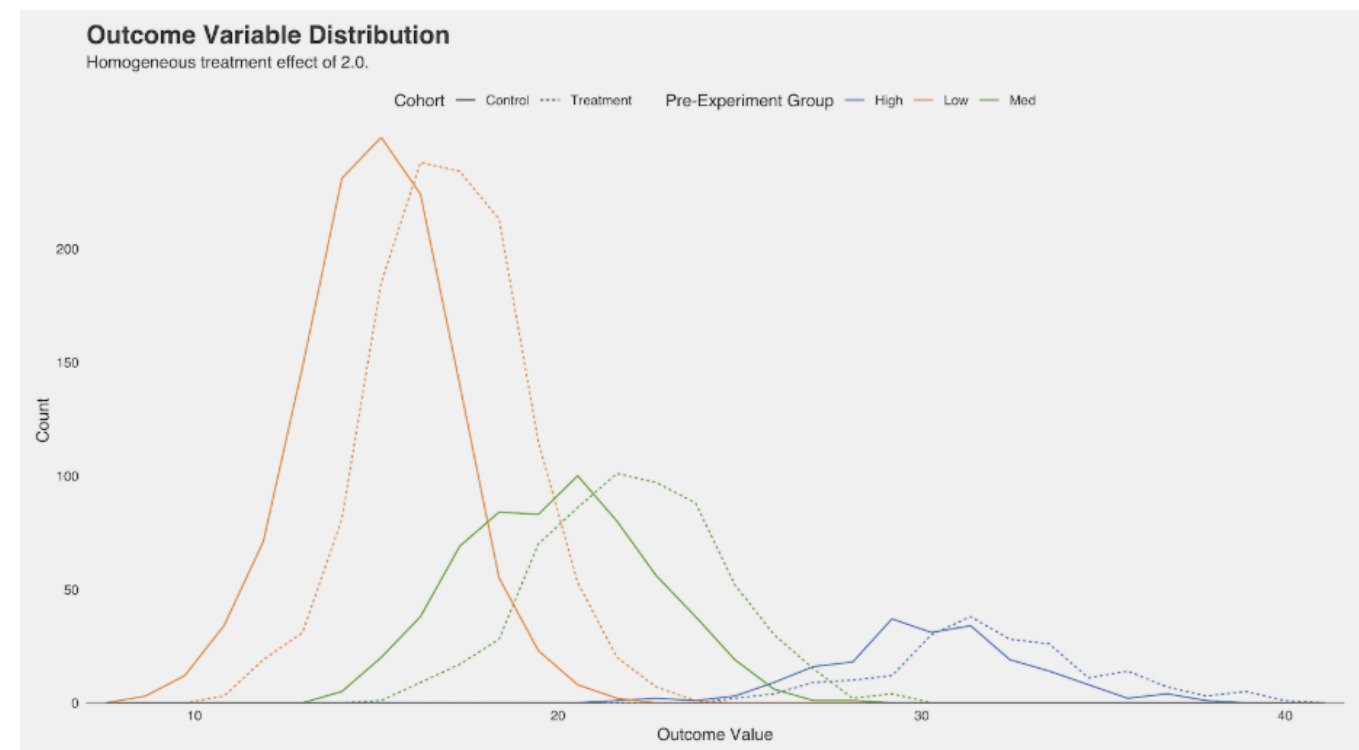
1. Age Groups
2. Genders
3. #visits in the last week
4. \$ spending in the last week
5. d(purchase) during the week before the experiment (high, low)

With the largest correlation  
with post-experiment OEC  
to remove the largest  
between-strata variance

# How to Divide Population into Strata

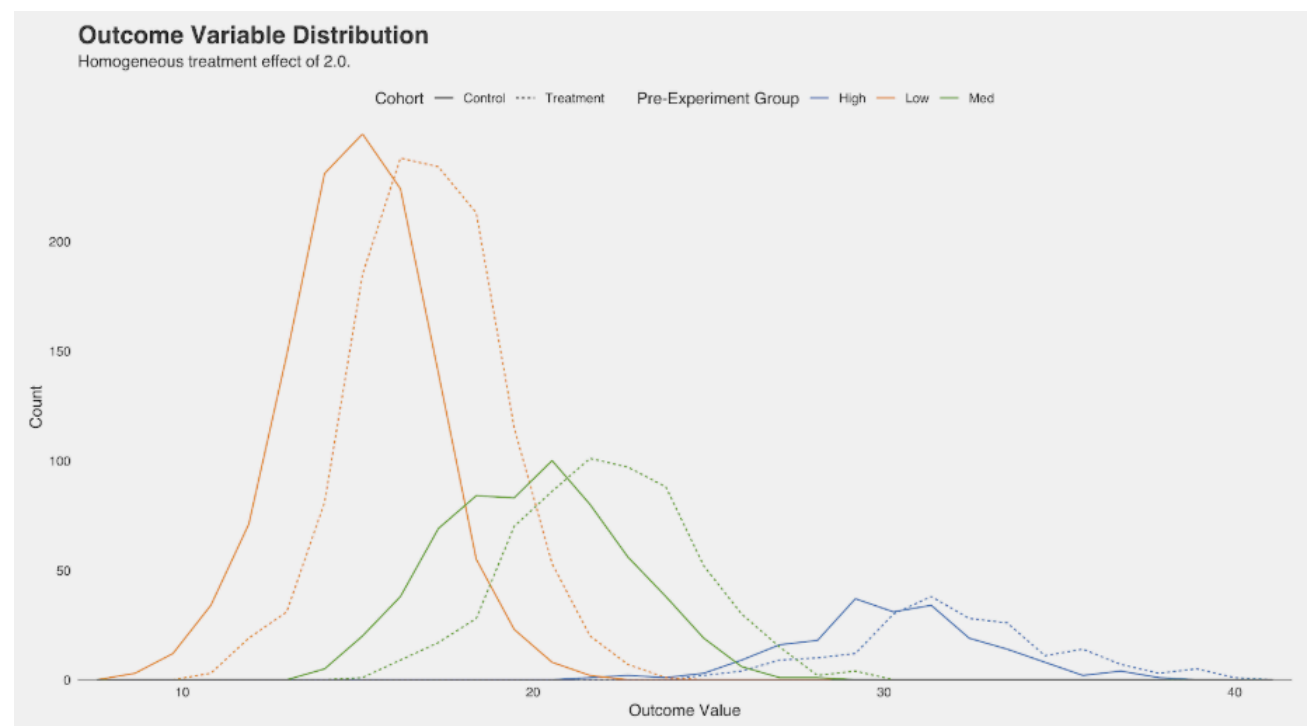
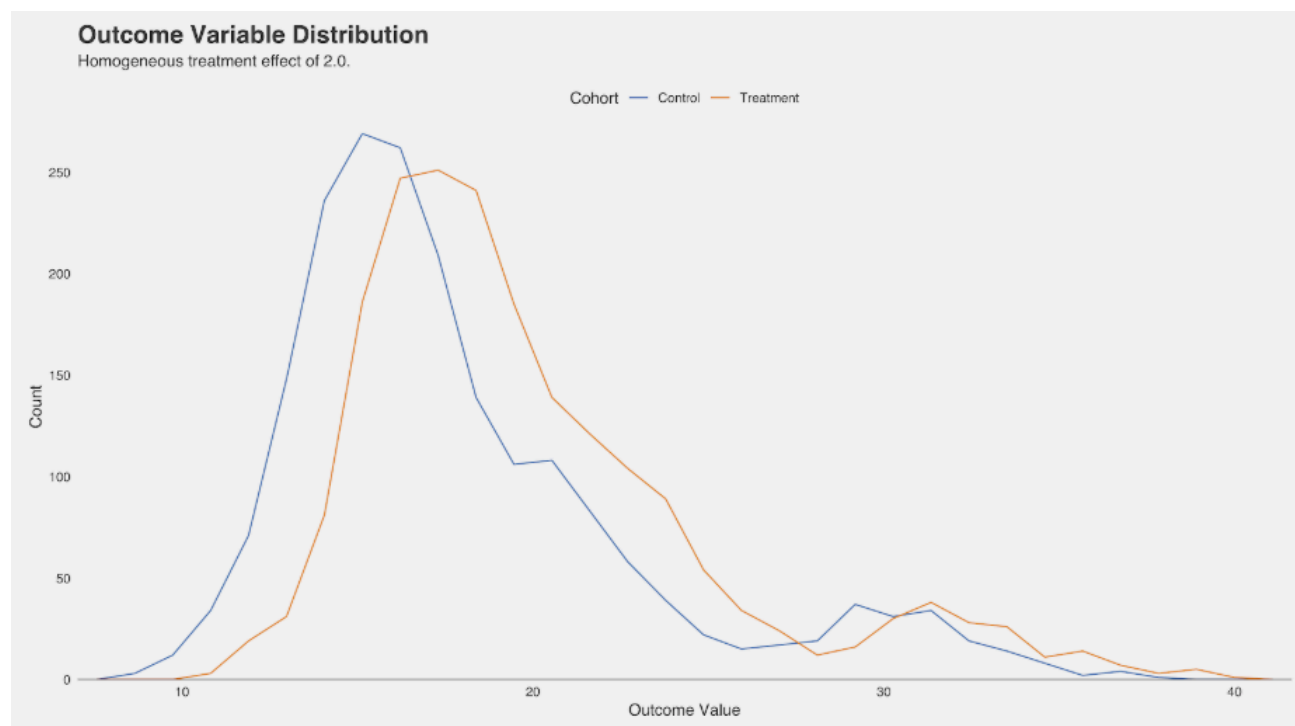
- Categorize **population** into different Strata
  - $n_k = p_k n$
- Pre-experiment OEC is a good choice and correlates well with OEC
- If OEC is a continuous variable
  - Transform pre-experiment OEC into a categorical variable
  - e.g., High, Middle, Low
  - Based on A/A test or Historical Data

Proportion of the stratum in the population



# Stratification

- For example, divide into three strata based on pre-experiment Y (OEC):
  - High, Middle, Low
- Estimate three treatment effects and then combine them together (weighted average)



# How to Divide Population into Strata

- When pre-experiment data on OEC is not available
  - A new product
  - Experiment with new users
- Choose the one, most likely highly correlated with the OEC
  - Based on experience and data
  - e.g., two UIs for shopping cart and OEC is purchase rate
    - Age or Gender

# Stratification

1. Pre-stratify the sample,  $n_k = p_k n$
2. Randomize completely within each stratum
3. Combine the Treatment Effects with different weights

$$\bar{Y}^s = \sum_{k=1}^K p_k \bar{Y}_k$$

$$\Delta^s = \bar{Y}_1^s - \bar{Y}_0^s$$

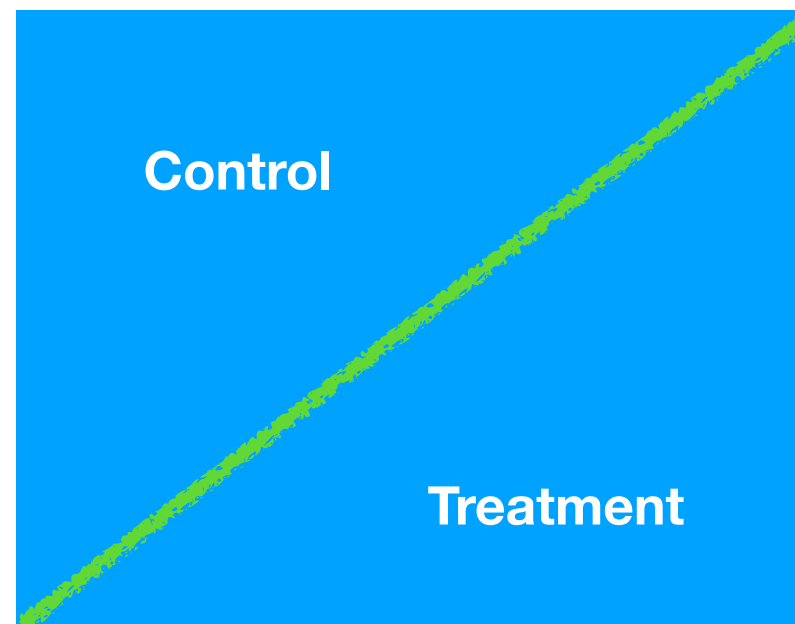
4. Calculate the variance

$$Var(\bar{Y}^s) = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2$$

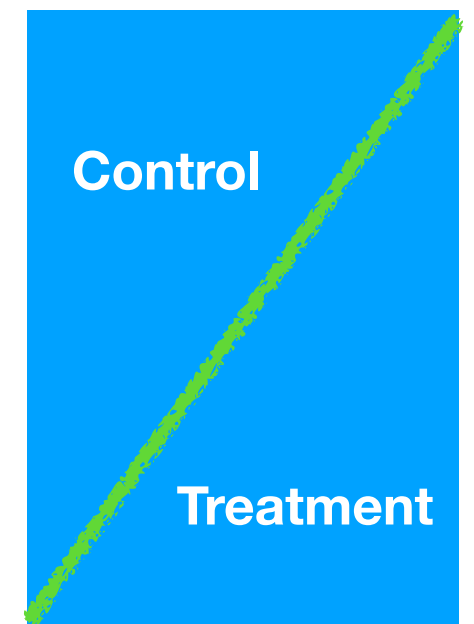
$$Var(\Delta^s) = Var(\bar{Y}_1^s) + Var(\bar{Y}_0^s)$$

5. Calculate t (z) stat

$$t = \frac{\Delta^s}{\sqrt{Var(\Delta^s)}}$$



High Pre-OEC



Low Pre-OEC

# Stratification

- Assume there are  $k$  strata:

$$\bar{Y}^s = \sum_{k=1}^K p_k \bar{Y}_k$$

$p_k$ : proportion of the population in the  $k$ -th stratum

$n_k$  : number of users from the  $k$ -th stratum

$$n_k = p_k n$$

$$E(\bar{Y}^s) = \sum_{k=1}^K p_k E(\bar{Y}_k) = \sum_{k=1}^K p_k \mu_k = \mu$$

- $\Delta^s$  under stratified sampling is an unbiased estimate of  $\delta$

$$\Delta^s = \bar{Y}_1^s - \bar{Y}_0^s$$

$$E(\Delta^s) = E(\bar{Y}_1^s - \bar{Y}_0^s) = \mu_1 - \mu_0 = \delta$$

- However,  $\text{Var}(\Delta^s) < \text{Var}(\Delta)$

$$\text{Var}(Y) = \sum_{k=1}^K p_k \sigma_k^2 + \sum_{k=1}^K p_k (\mu_k - \mu)^2 = \text{Var}(Y^s) + \sum_{k=1}^K p_k (\mu_k - \mu)^2$$

$$\text{Var}(\bar{Y}^s) = \text{Var}(\bar{Y}) - \frac{1}{n} \sum_{k=1}^K p_k (\mu_k - \mu)^2$$

Remove Between-Strata Variance

$$\text{Var}(\Delta^s) = \text{Var}(\bar{Y}_1^s) + \text{Var}(\bar{Y}_0^s)$$

# Class Exercise

OEC is the \$Purchase/week

$$p(\text{male})=p(\text{female})=0.5$$

$$n=1000$$

Sample Characteristics:

- $n(\text{male})=n(\text{female})=0.5*1000=500$
- $\bar{Y}(\text{male})=20, \bar{Y}(\text{female})=50$
- $\text{Var}(\text{male})=20, \text{Var}(\text{female})=10$

A. What are the mean and variance for Y?

B. What are the stratification mean and its variance?



# Class Exercise

A. What are the mean and variance for the whole sample?

$$\bullet \quad \bar{Y}^s = \bar{Y} = \sum_{k=1}^K p_k \bar{Y}_k = 0.5 * 20 + 0.5 * 50 = 35$$

$$\bullet \quad Var(Y) = \sum_{k=1}^K p_k \sigma_k^2 + \sum_{k=1}^K p_k (\mu_k - \mu)^2 = Var(Y^s) + \sum_{k=1}^K p_k (\mu_k - \mu)^2$$

$$= 15 + 0.5 * (20 - 35)^2 + 0.5 * (50 - 35)^2$$

B. What are the stratification mean and its variance?

$$\bullet \quad \bar{Y}^s = \sum_{k=1}^K p_k \bar{Y}_k = 0.5 * 20 + 0.5 * 50 = 35$$

$$\bullet \quad Var(\bar{Y}^s) = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2 = 1/1000 * (0.5 * 20 + 0.5 * 10) = 0.015$$

# Case: Netflix's Implementation of Stratification

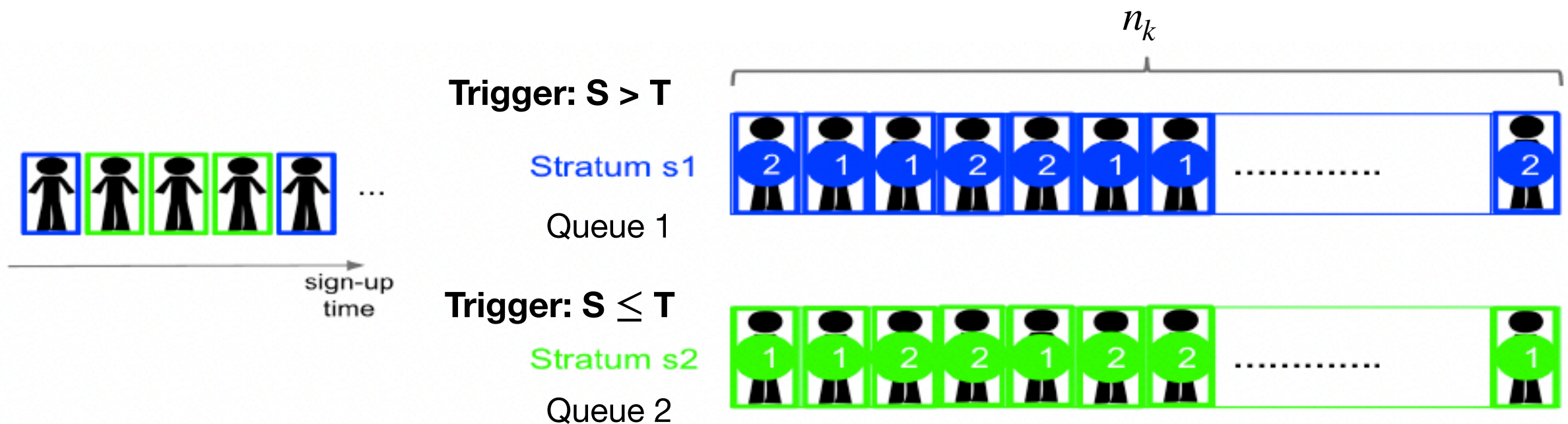
A. Randomize users into different variants for each stratum **in real-time**.

1. Define strata factor: pre-experiment streaming hours
  - Stratum (s) = high if factor > threshold (T)
  - Different T leads to different  $\rho$  = correlation (s, OEC)
    - $\rho > 0.5$
  - Only on existing users, not on new users **WHY?**
2. **Pre-define**  $n_k$  for each  $k$  (stratum) based on  $p_k$

# Case: Netflix's Implementation of Stratification

## B. Rely on a Queue System for (Random) Assignment

1. Assign users into different queues based on their strata factor (e.g., pre-OEC) - **trigger conditions**



# Case: Netflix's Implementation of Stratification

## B. Rely on a Queue System for (Random) Assignment

1. Assign users into a different queue based on their strata factor (e.g., pre-OEC) - a trigger
2. A queue consists of many 100-slot segments
3. Randomly assign users into different variants within each queue
  - A. Assign a number to each cell [1, 100]
  - B. Shuffle the numbers for the cells
  - C. Map 1 (control) to the cell with the number [1, 50], and 2 (treatment) with [51, 100]

(a)

1	2	3	4	5	6	.	.	.	.	.	.	.	100
---	---	---	---	---	---	---	---	---	---	---	---	---	-----

(b)

25	57	9	12	95	64	.	.	.	.	.	.	.	43
----	----	---	----	----	----	---	---	---	---	---	---	---	----

(c)

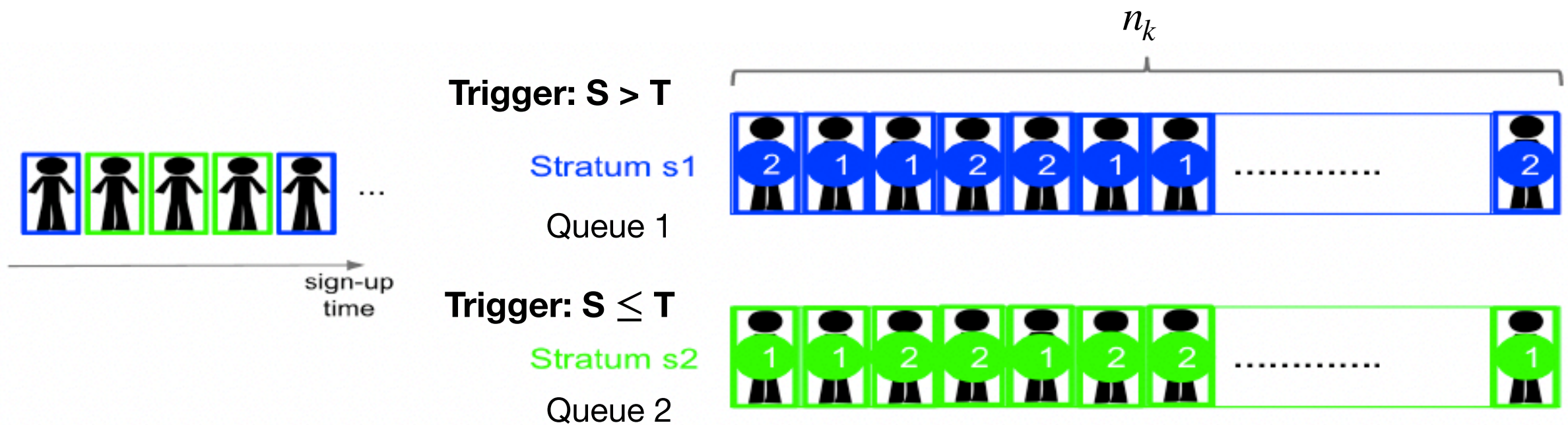
1	2	1	1	2	2	.	.	.	.	.	.	.	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---



# Case: Netflix's Implementation of Stratification

## B. Rely on a Queue System for (Random) Assignment

1. Assign users into different queue based on their strata factor (e.g., pre-OEC) - a trigger
2. Randomly assign users into different variants within each queue



# Post Stratification

- A popular post-assignment variance reduction technique
- Assume simple random assignment but uses the estimate of stratification:

$$\bar{Y}_{post}^s = \sum_{k=1}^K p_k \bar{Y}_k$$

$p_k$ : proportion of the **population** in the k-th stratum

$n_k$  : number of users from the k-th stratum

$$n_k = p_k n$$

$$E(\bar{Y}_{post}^s) = \sum_{k=1}^K p_k E(\bar{Y}_k) = \sum_{k=1}^K p_k \mu_k = \mu$$

# Post Stratification

1. Randomly assign users into control and treatment groups.

2. Collect data on OEC and S (strata factor)

3. Calculate  $\bar{Y}_k$  for each stratum.

4. Calculate  $\bar{Y}_{post}^s = \sum_{k=1}^K p_k \bar{Y}_k$

- $p_k$  is predefined according to “population” data
- $\bar{Y}_k$  is sample mean for each stratum k

5. Calculate  $\Delta_{post}^s = \bar{Y}_{post1}^s - \bar{Y}_{post0}^s$

6. Calculate  $\text{Var}(\Delta_{post}^s)$

$$\text{Var}(\bar{Y}_{post}^s) = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2 + \frac{1}{n^2} \sum_{k=1}^K (1 - p_k) \sigma_k^2 + o\left(\frac{1}{n^2}\right)$$

Larger than  $\text{Var}(\bar{Y}^s)$

$$\text{Var}(\Delta_{post}^s) = \text{Var}(\bar{Y}_{post1}^s) + \text{Var}(\bar{Y}_{post0}^s)$$

# Post Stratification

- However, variance reduction is smaller than stratification at assignment
  - $n_k^{post} \neq p_k n$
- $var(\bar{Y}^s) \leq var(\bar{Y}_{post}^s) \leq var(\bar{Y})$
- When  $n$  is very large:
  - $var(\bar{Y}^s) \approx var(\bar{Y}_{post}^s) \leq var(\bar{Y})$
- Post stratification is less costly to implement.



# Example

OEC is the \$Purchase/week

$$p(\text{male})=p(\text{female})=0.5$$

$$n=1000$$

Sample Characteristics:

- $n(\text{male})=400$
- $n(\text{female})=600$
- $\bar{Y}(\text{male})=20, \bar{Y}(\text{female})=50$
- $\text{Var}(\text{male})=20, \text{Var}(\text{female})=10$

Different from  $p_k$


Stratification Mean and Variance (Mean)

$$\bar{Y}^s = \sum_{k=1}^K p_k \bar{Y}_k = 0.5 * 20 + 0.5 * 50 = 35$$

$$\text{Var}(\bar{Y}_{post}^s) = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2 + \frac{1}{n^2} \sum_{k=1}^K (1 - p_k) \sigma_k^2 + o\left(\frac{1}{n^2}\right) = (0.5 * 20 + 0.5 * 10) / 1000 + (0.5 * 20 + 0.5 * 10) / 1000^2 + o\left(\frac{1}{n^2}\right)$$

# CUPED

## CUPED: Controlled Experiments by Utilizing Pre-Experiment Data (Deng, Xu, Kohavi, & Walker, 2013)

- Remove variance in a metric that can be accounted for by pre-experiment information.
- Control Variates: Pre-experiment information  $X$ .
- $Y^{cuped} = Y - \theta X$   Remove the variance caused by  $X$
- $\Delta^{cuped} = \Delta = m_1 - m_0 = \bar{Y}_1 - \bar{Y}_0 = \bar{Y}_1^{cuped} - \bar{Y}_0^{cuped}$

# How to Choose Control Variables

- $Var(Y^{cuped}) = Var(Y) + \theta^2 Var(X) - 2\theta Cov(X, Y)$
- To Minimize ( $Var(Y^{cuped})$ ),  $\theta = Cov(X, Y)/Var(X)$
- $Var(Y^{cuped})_{min} = Var(Y)(1 - \rho^2)$ 
  - $\rho = \text{correlation}(X, Y)$
- Choose the X with the largest correlation with Y (e.g., post-experiment OEC).
- Empirically, pre-experiment OEC correlates well with Y.
- e.g., OEC is purchase amount/user, X is pre-experiment purchase amount

# How to Choose Control Variables

“Across a large class of metrics, our results consistently showed that using the same variable from the pre- experiment period as the covariate tends to give the best variance reduction. ” (Deng, Xu, Kohavi, & Walker, 2013)

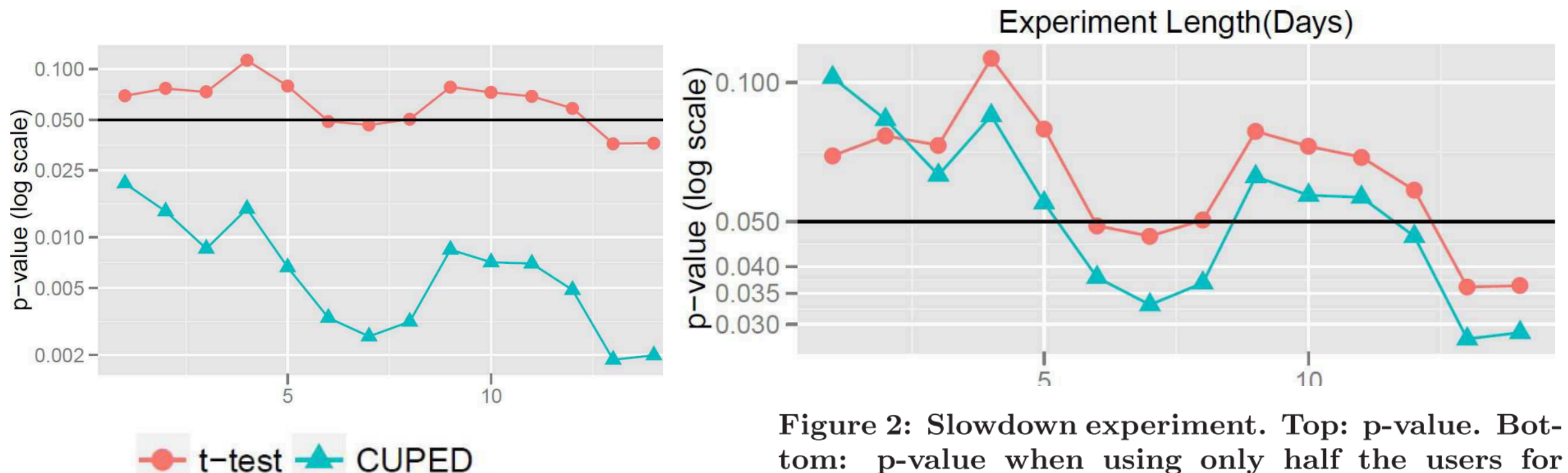


Figure 2: Slowdown experiment. Top: p-value. Bottom: p-value when using only half the users for CUPED.

# Class Exercise

- Use exp\_data\_3.csv
- Find the Treatment Effects on Clicking Ads using CUPED
  - $X$ : pre\_click
  - $Y^{cuped} = Y - \theta X$
  - $\theta = Cov(X, Y) / Var(X)$
- Find the variance reduction by CUPED
  - Compare  $Var(\Delta^{cuped})$  vs.  $Var(\Delta)$
  - p value

# Class Exercise

```
var_x=np.var(df.pre_click, ddof=1)
#np.cov returns a var-cov metrix
cov_xy = np.cov(df.pre_click,df.click, ddof=1)[0][1]
theta = cov_xy/var_x
df['theta']=theta
df['click_cuped']=df.click - df.pre_click*theta
```

## With CUPED

```
d_0 = df[df['treat'] == 0]['click_cuped']
d_1 = df[df['treat'] == 1]['click_cuped']
diff = np.mean(d_1) - np.mean(d_0)
print(diff)

cm = sms.CompareMeans(sms.DescrStatsW(d_1), sms.DescrStatsW(d_0))
ttest = cm.ttest_ind(alternative = 'two-sided', usevar = 'unequal')
se = cm.std_meandiff_separatevar
print(se,ttest)
```

# Class Exercise

## Without CUPED

```
d_0 = df[df['treat'] == 0]['click']
d_1 = df[df['treat'] == 1]['click']
diff = np.mean(d_1) - np.mean(d_0)
print(diff)

cm = sms.CompareMeans(sms.DescrStatsW(d_1), sms.DescrStatsW(d_0))
ttest = cm.ttest_ind(alternative = 'two-sided', usevar = 'unequal')
se = cm.std_meandiff_separatevar
print(se, ttest)
```

# Regression with Control Variables

- Remove the variances explained by X (covariates, e.g., age, gender, pre-experiment behaviors)
- X can be continuous variables
- Stratification: Strata Factor has to be categorical variable

## **Regression with Control Variates**

$$Y_i = \beta_0 + \beta_1 \cdot T_i + X_i\theta + \epsilon_i$$

$T_i = \{0,1\}$ , Control or Treatment

$Y_i$  is the OEC (or other metrics)

X are control variables



# Regression with Control Variables

$$Y_i = \beta_0 + \beta_1 \cdot T_i + X_i\theta + \epsilon_i$$

- Linear regression gives a consistent estimator for the average treatment effect
  - $\beta_1$  represents the Average Treatment Effects
- Reduces variance by controlling for X
- BUT, it makes assumption:
  - the conditional expectation of the outcome metric is linear in the treatment assignment and covariates.

# Class Exercise

- Use exp\_data\_3.csv
- 1. Find the Treatment Effects of Clicking Ads using OLS:
  - $Y_i = \beta_0 + \beta_1 \cdot T_i + \epsilon_i$
- 2. Find the correlations between X and Y.
  - X: gender and pre-click (i.e., whether clicking ads during the week before the experiment)
- 3. Adding Control Variables **separately and then sequentially**
  - $Y_i = \beta_0 + \beta_1 \cdot T_i + X_i\theta + \epsilon_i$
- 4. What changes and What does not change after adding control variables?
- 5. Find the maximum variance reduction you can get by adding the control variables.

# Class Exercise

```
import statsmodels.formula.api as smf
mod = smf.ols(formula='click ~ treat +
pre_click', data=df)
res = mod.fit()
print(res.summary())
```

# Wrap-up

1. A/B testing Terminology and Overview
  2. Statistics behind A/B testing
    1. Statistical tests (t, z, chi-square)
    2. Confidence intervals
    3. Type I error & Multiple Testing
    4. Type II error & Power Analysis
    5. Regression
  3. Internal & External Validity
    1. SUTVA (network interferences)
    2. Survivorship bias
    3. Sanity Checks (SRM, Randomization checks, A/A tests)
    4. Heterogeneous Treatment Effects
  4. Improve Sensitivity
    1. Estimate  $\sigma^2$ : ratio metrics (lift), Clustered SE (correlated observations)
    2. Increase N (pooled control group, split sample)
    3. Increase effect size (Triggering Experiments)
    4. Reduce variance (transform matrix and interleaving design)
    5. Stratification (post and at assignment)
    6. Regression with controls, CUPED
    7. Paired Design, Block Design
- **Compare the means (lift, median, etc) between treatment and control**
  - **Interpret the results considering type I and II errors**
  - **Two principles to be considered during the whole process of experiments**
  - **Need to guarantee internal validity**
  - **Consider external validity when generalizing the results**
  - **Improve sensitivity means using the smaller sample to achieve larger power**
  - **Always a desire to improve the power given a sample size**

# A/B Testing

- A/B testing is relatively new, particularly in China.
- My goal is to teach you how to approach and solve real-world problems effectively.
- It's important to recognize that there isn't a one-size-fits-all solution; rather, you should apply what you've learned to a range of situations.
- Deepening your understanding will take time and experience.
- Currently, there are fewer experts in A/B testing and causal inference compared to those specializing in machine learning and predictions within the industry.
- The shift towards data-driven decision-making—where machines are increasingly making decisions traditionally made by humans—is underway.
- This course is designed to equip you to be an active participant in this evolving landscape.

- Feel free to email me or make an appointment with me if I am of any help.
  - [shanh@hku.hk](mailto:shanh@hku.hk)
  - KKL 1229
  - <https://www.shanhhuang.com/>
- Thank you very much!

