BITTIGER

DS306 数据科学面试 - AB Test专题

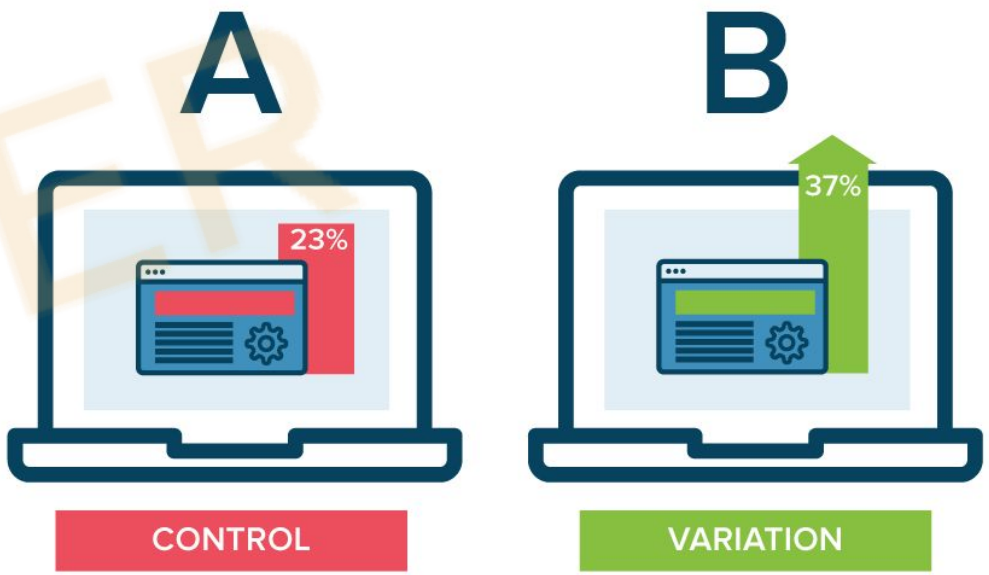
你的专业背景？

Do you know how to answer the following questions?

- Survey showed teenager users are less engaged on your app after their parents join. What to do?
- You have 1M budget to spend on holiday campaigns. Possible investments include mailed ads / emails / display ads / search engine / social media ads. How would you optimize the budget?
- An engineer suggest promoting new sellers on your website to boost seller growth. But another engineer is worried about it hurting overall sales. How would you make a decision?

Outline

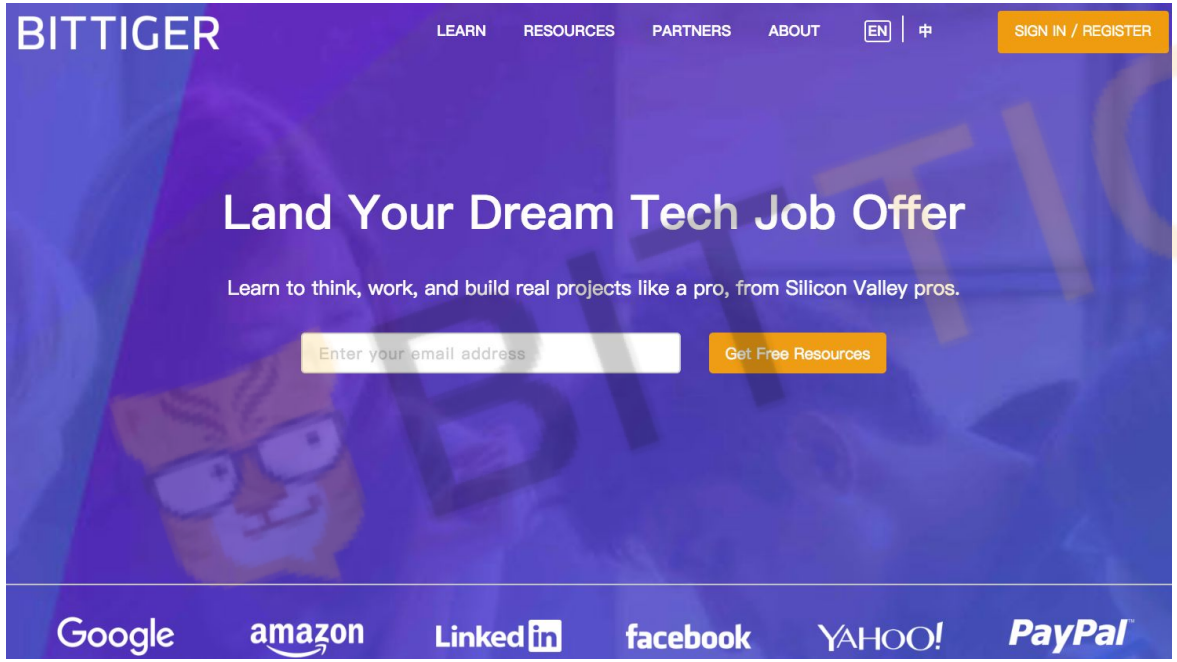
- Introduction
- Statistical Foundations
- A/B test process
  - Design of Experiment
  - Result Measurement
- Advanced A/B test topics
- A/B Test Interviews

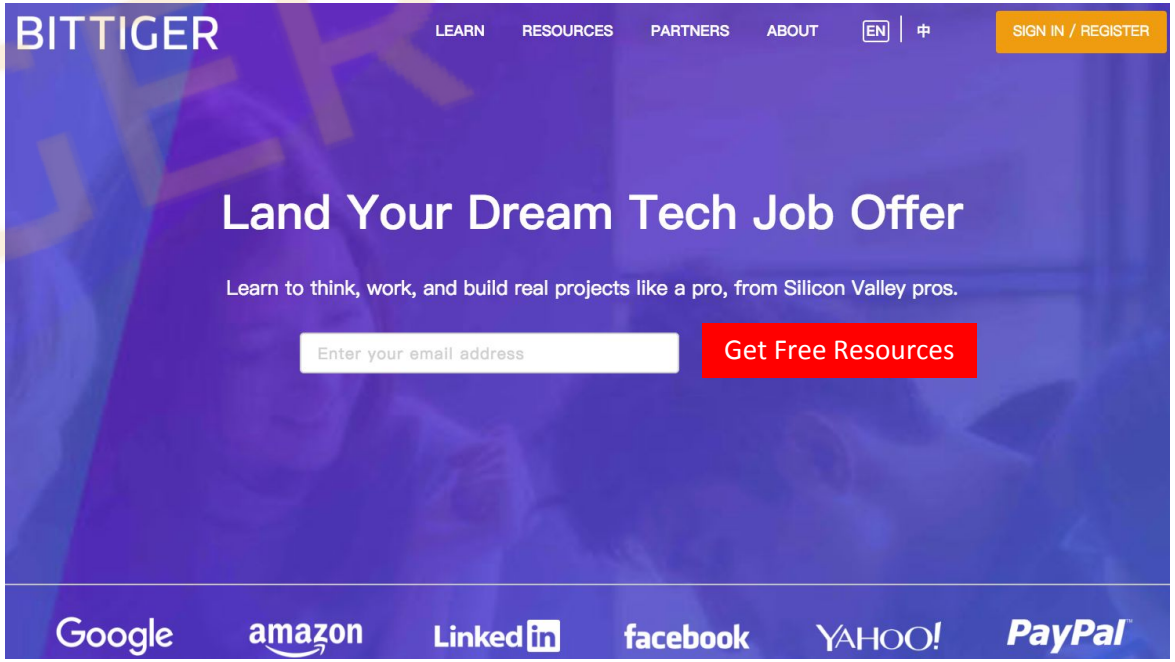


Introduction

What is A/B Test?

A/B testing is general framework of hypothesis testing between two groups



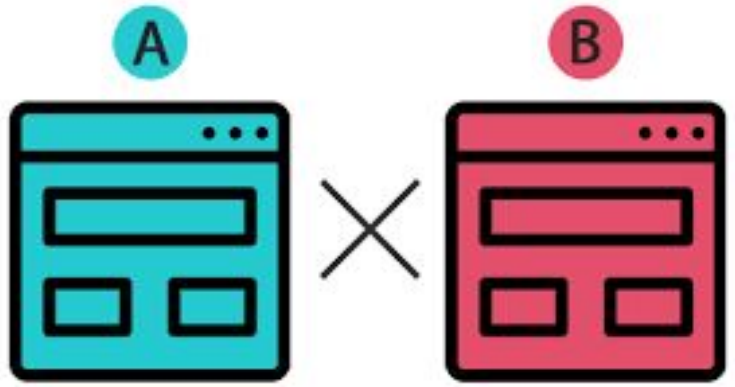


Introduction

Why do we need A/B Test?

The goal is to

- Establish causal relationship between actions and results
- Measure impact solely from the change



Introduction

Where is A/B Test used?

Widely used in high tech industry. Major use cases

- Product iteration

  Examples:

  - Front End: change UI design, user flow, add new features
  - Algorithm Enhancement: recommendation system, search ranking, ads display
  - Operations: define coupon value, promotion program

- Marketing optimization

  Examples

  - Search engine optimization (SEO)
  - Campaign performance measurement

In other industries, there are other forms of experiments / tests (e.g. clinical experiments in biostatistics). We will focus on A/B test in tech industry for this class

Introduction

What's data scientist's role in A/B test?

Product Team

- Design of Experiment
- Result Measurement / Analysis
- Product Insights
- Launch Decisions

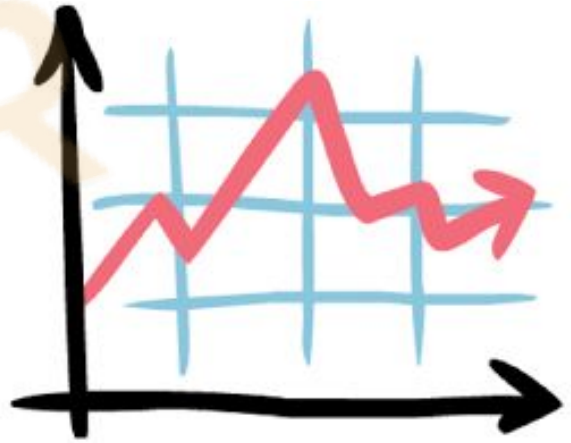
Platform Team

- Design a A/B testing platform
- Define methodology

Statistical Foundations

Statistical Foundations Outline

- Normal Distribution
- Central Limit Theorem
- Correlation != Causation
- Hypothesis Testing
  - T test Z test
  - P value
  - Two sample / One sample / Paired



Normal Distribution

- Normal distribution

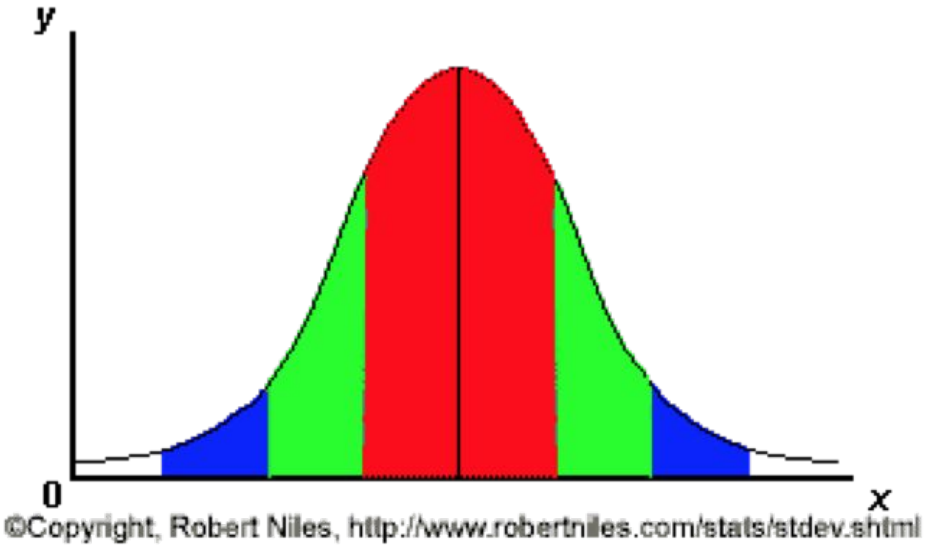
$$P(X = x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

$$E[X] = \mu \text{ and } Var(X) = \sigma^2$$

- Standard normal distribution (Z) $\mu = 0$ and $\sigma = 1$

$$X = \mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$$

- Beauty of normal curve (6σ)
  - [μ - 3σ, μ + 3σ] covers 99.7%
  - [μ - 2σ, μ + 2σ] covers 95%
  - [μ - σ, μ + σ] covers 68%



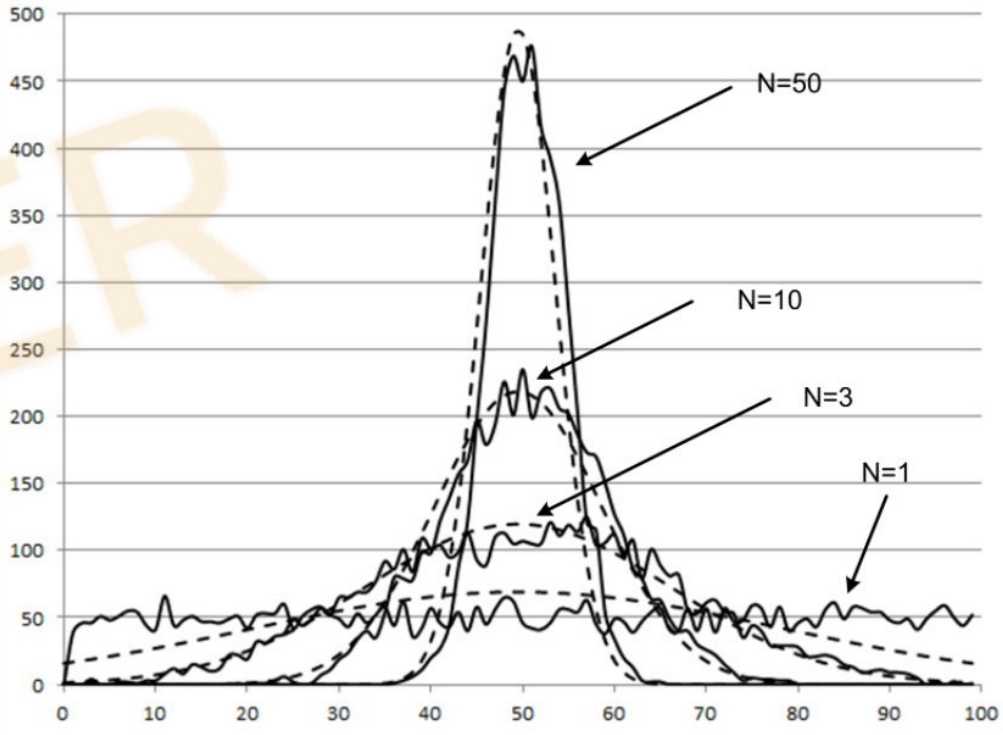
Central limit theorem

- $X_1, X_2 \ldots X_n \ldots$ are independent, identically - distributed (IID) random variables, X_i has finite mean μ and variance σ^2

$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

(replacing σ by sample standard deviation, CLT still holds)

- Application
  - Binomial distribution



CLT - Interview Quiz

User CTR of your website is p

You sampled 1000 users from your website. What is the sampling distribution of the sample mean?

$$\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

When CLT doesn't hold?

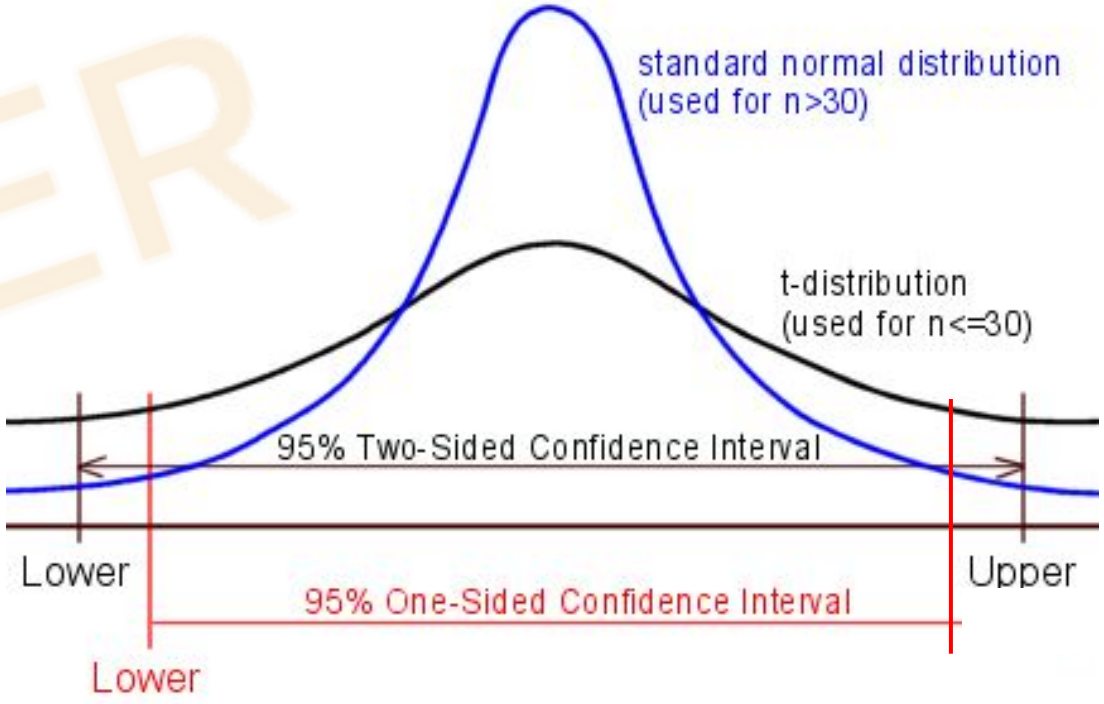
- Normal distribution -> T distribution when N < 30
  - T distribution has only one parameter: degree of freedom (df = N-1)
  - Approximate normal as df increases
  - CI under normal distribution

$$Mean_{estimate} \pm z_{1-\alpha/2} * StdErr_{estimate}$$

  - CI under t distribution

$$Mean_{estimate} \pm t_{n-1} * StdErr_{estimate}$$

  - z or t?



Causal inference

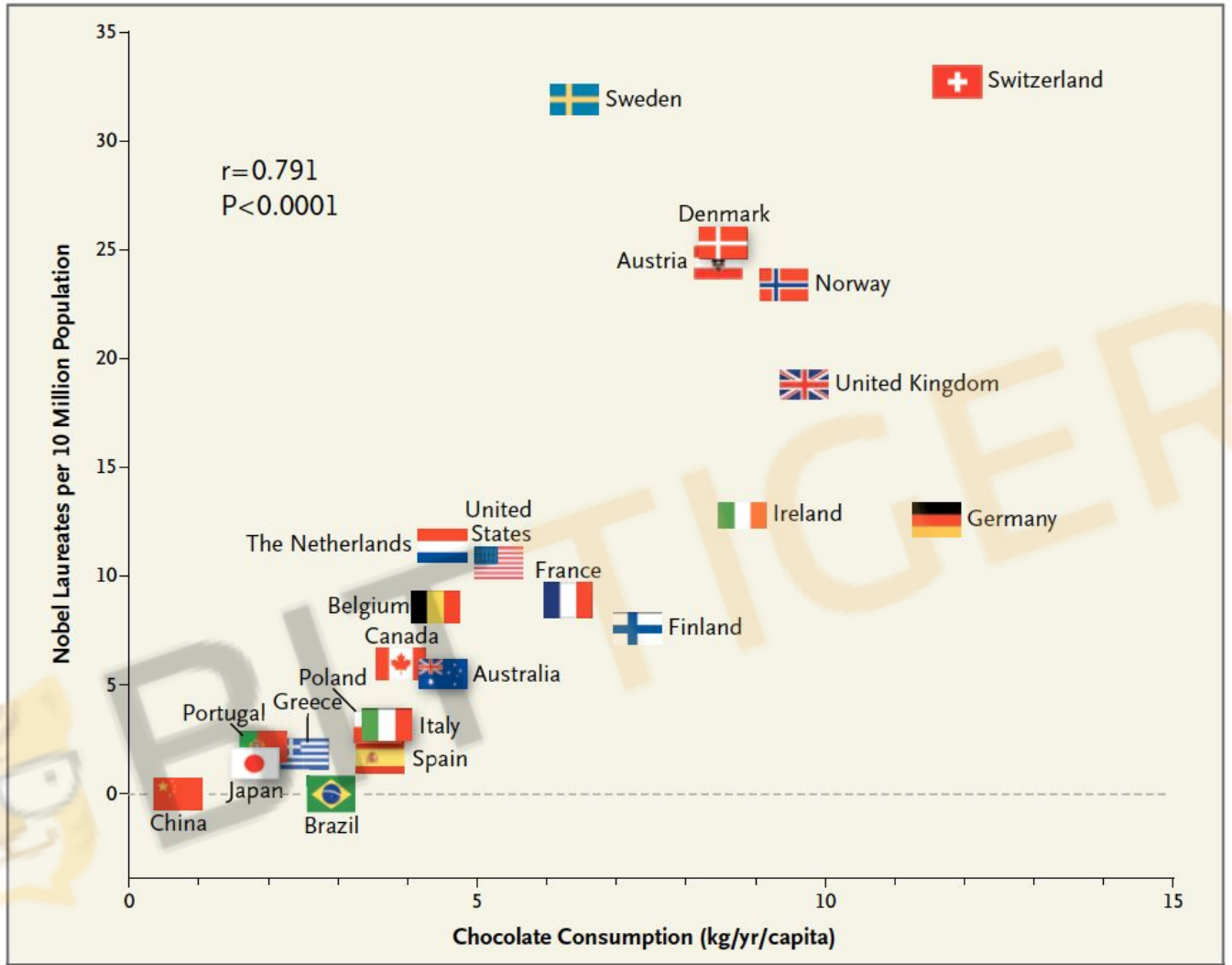


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

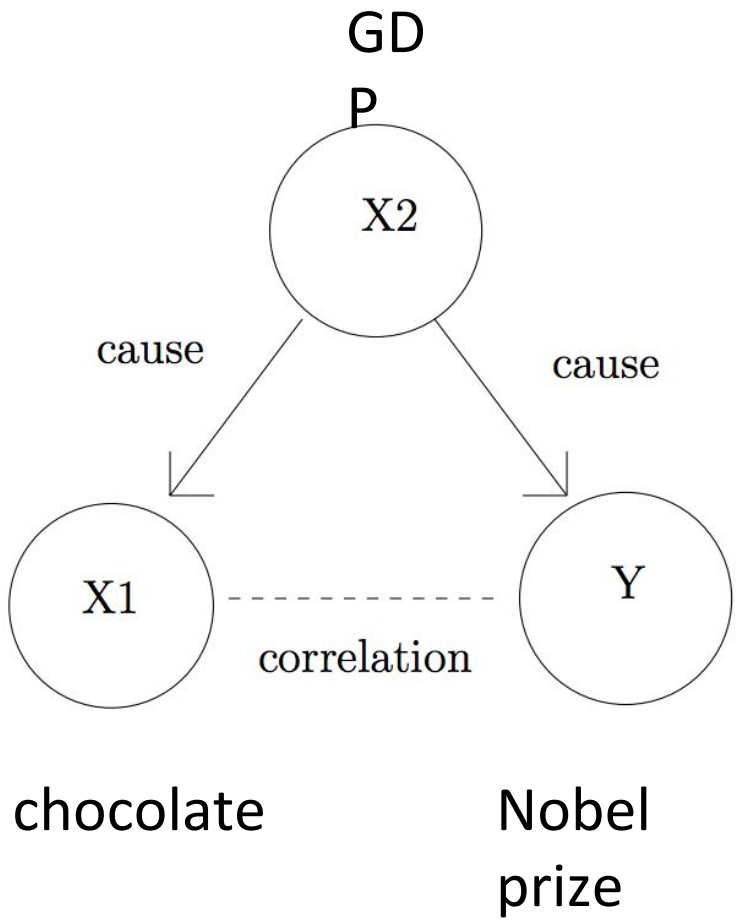
Strong correlation between chocolate consumption & Nobel laureates.

Does eating chocolate making people more likely to win Nobel Laureates?

Correlation is not causality

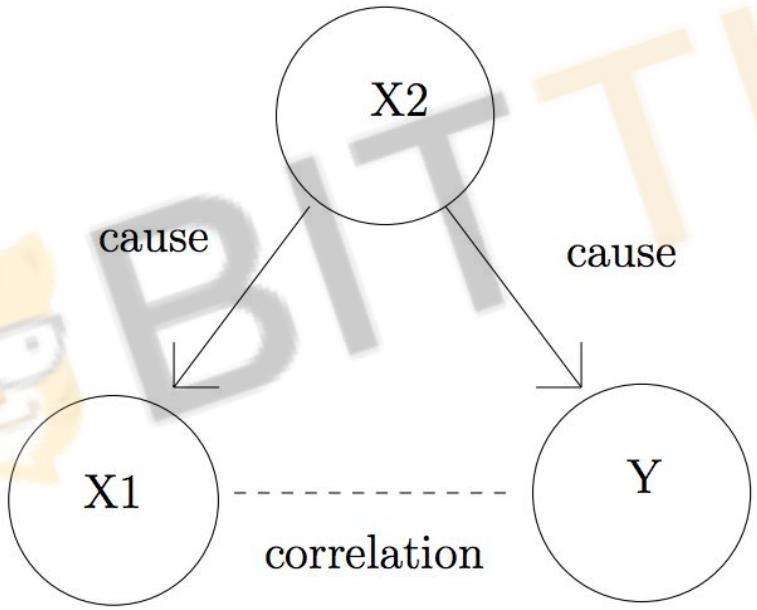
- Quote from chairman of Nobel chemistry committee
  - “Chocolate is a luxury. Wealthy individuals are more likely to be able to afford it.

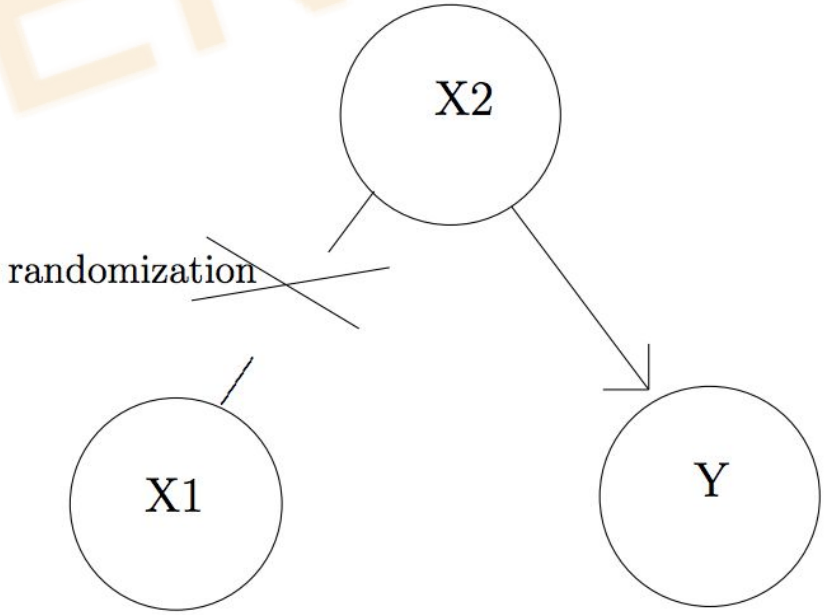
    Education is also a luxury. Poor people can’t afford to go to college for 10 years to get a PhD in chemistry. But you can’t win the Nobel prize in chemistry unless you’re a chemist. “
- Common factor
  - “GDP or wealth of a country will be correlated both with chocolate eating and with Nobel prizes. ”



Observational study v.s. Randomized experiment

- Observational studies can suggest good experiments to run, but can't definitively show causality.
- Randomization can eliminate correlation between X_1 and Y due to a different cause X_2(confounder).





Design randomized experiments

- Define the causal relationship to be explored, X -> Y
  - New UI decreases user interaction
- Define metric (Y)
  - Number of posts per user per day
- Design randomized experiments (A/B test)
  - Two groups of users, comparable
  - control group: old UI, experiment group: new UI
- Collect data and conduct **hypothesis testing**
  - Compare the metrics using two sample t test
- Draw conclusion

Hypothesis testing

- Definition
  - Use sample of data to test an assumption regarding a population parameter, which could be
    - A population mean μ
    - The difference in two population means, $\mu 1 - \mu 2$
    - A population variance
    - The ratio of two population variances
    - A population proportion p
    - The difference in two population proportions, $p1 - p2$

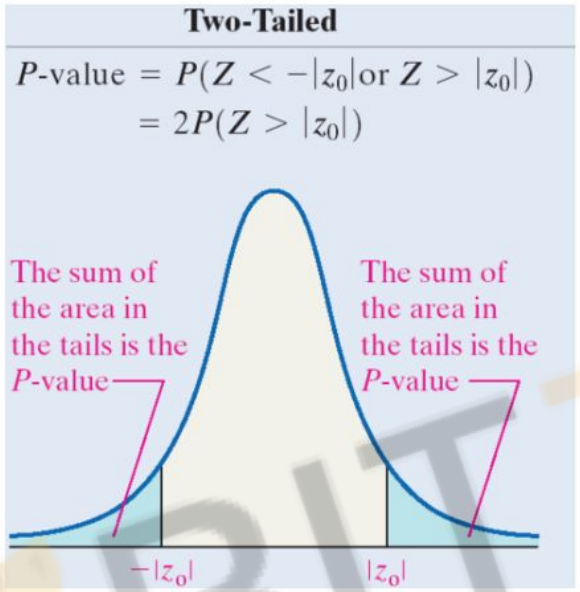
Hypothesis testing (cont'd)

- Definition
  - Two opposing hypotheses about a population
    - Null hypothesis, H_0 , is usually the hypothesis that sample observations result purely from chance.
    - Alternative hypothesis, H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

Different types of alternative hypothesis

- Two tailed v.s. one tailed

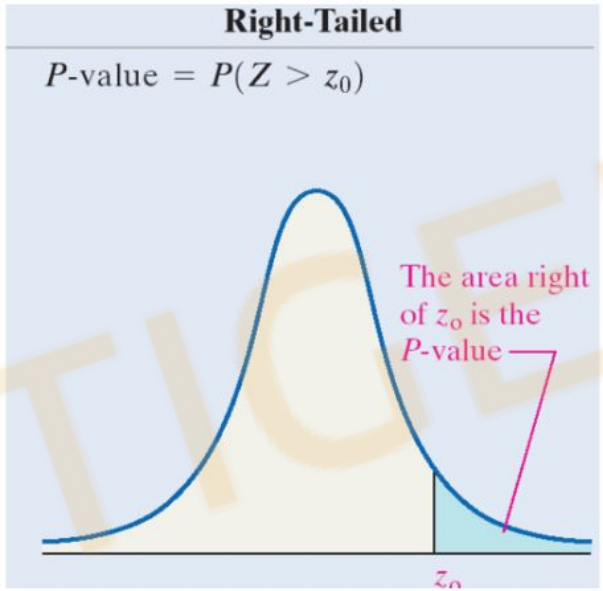
Two-tail



$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

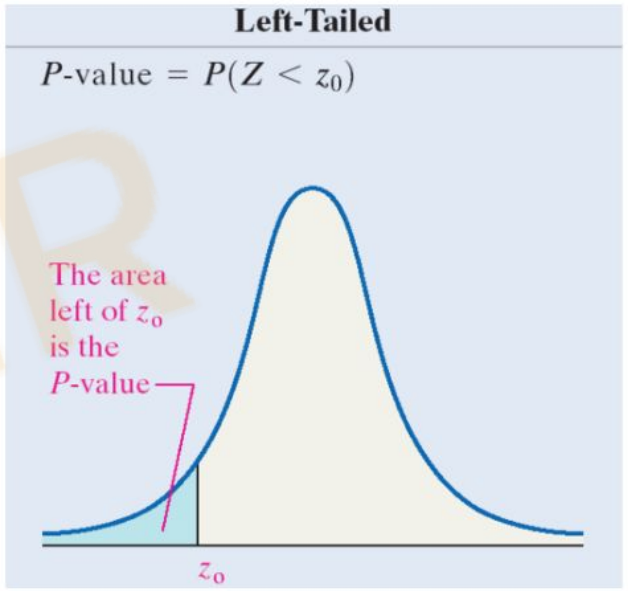
Right Tail



$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

Left Tail



$H_0 : \mu = \mu_0$

$H_1 : \mu < \mu_0$

- When to use one sided test? Reading

How to construct hypothesis testing

Scenario: flip coin 50 times and see 40 heads. Is this a fair coin?

1 $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

Given sample data

p is population statistic of our interest

2 $H_0 : p = \frac{1}{2}$ $H_1 : p \neq \frac{1}{2}.$

Null hypothesis

Alternative hypothesis

3 How to decide if we should reject or not reject null hypothesis?

How to decide reject or not reject?

P_value: Assuming null hypothesis is true, what's the probability of observing a as extreme or more extreme test statistics as the observed case

Example:

flip coin 10 times, see 7 heads

P_value = Observe (>=7 heads | flip fair coin 10 times)

Flip coin 100 times, see 77 heads

P_value = Observe (>=77 heads | flip fair coin 100 times)

Two ways to decide

- "Critical value" approach, compare z with critical value
- "P value" approach, compare p value with threshold (type I error)

Z test v.s. T test

Hypothesis test for the population mean

- If population variance σ^2 is known and n is large, z test
- If the population variance σ^2 is unknown (most of the time), t test

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Reject H0 if z > z*

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

Reject H0 if t > t*

One sample v.s. Two sample tests

- One sample test
  - One population, compare test statistic, e.g, sample mean, with a known number
- Two sample test
  - Two populations, compare two population means
  - Paired test
    - Two dependent groups, for example, same group been measured at two different times.
      - Essentially one sample test
  - Unpaired test
    - Two independent groups, may have different sample sizes.

Two sample t-test

- Compares the means of the two groups of data
  - X_1 random sample from $N(\mu_1, \sigma_1^2)$

    X_2 random sample from $N(\mu_2, \sigma_2^2)$
  - $H_0 : \mu_2 = \mu_1 \quad H_a : \mu_2 != \mu_1$ (other H_a?)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{Var(\bar{x}_1 - \bar{x}_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{Var(\bar{x}_1) + Var(\bar{x}_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Student t test v.s. Welch t test

- If population variance from two samples are equal, use pooled variance (student t test)

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \; df = n_1 + n_2 - 2$$

- If population variance from two samples are not equal, use unpooled variance (Welch t test)

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{(n_1-1)\cdot(n_2-1)}{(n_2-1)C^2+(1-C)^2(n_1-1)} \qquad C = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

A simplified way $df = \min(\mathrm{n1} - 1, \mathrm{n2} - 1)$

Hypothesis Test - Interview Quiz

- Test a coin is fair. What test? What is null hypothesis?
- Test two group of users have same CTR (click through rate). What test? What is d.f. (degree of freedom)?
- Test two group of users on your website have the same mean spending. What test? What is d.f.?

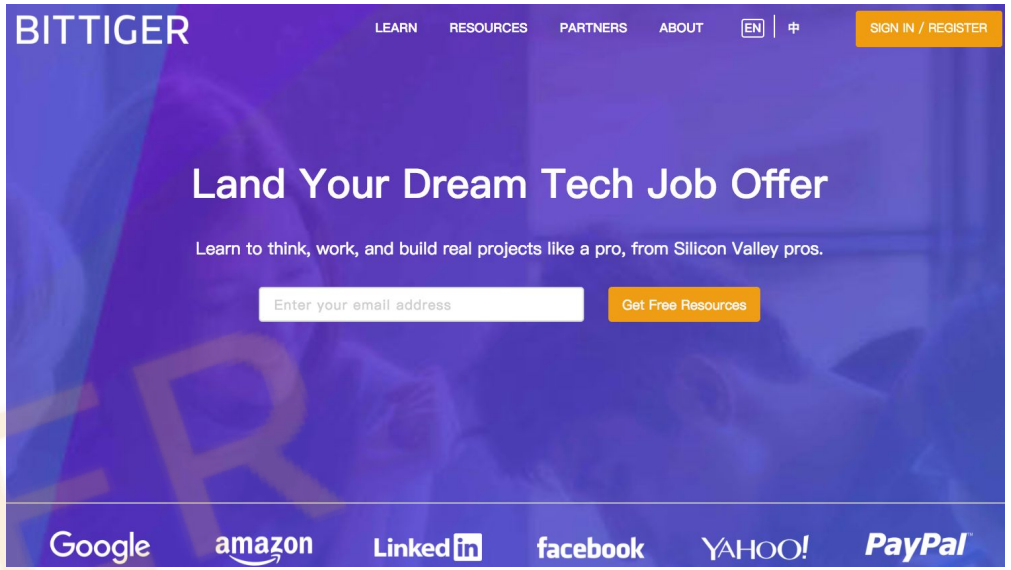
Assumptions of t test

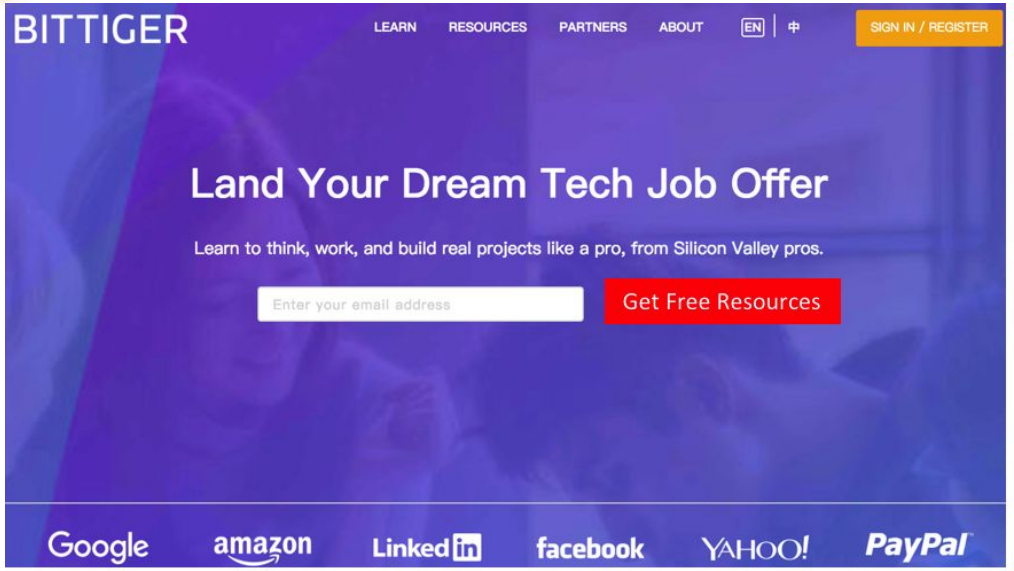
- Student t test
  - Normality
  - Independence
  - Equal variance (two sample test)
- What if normality is violated
  - Just do it (CLT)
  - Transformation
  - Other nonparametric methods

Process of A/B Testing

Process of A/B Test

- Design
  - Understand Problem & Objectives
  - Come Up with Hypothesis
  - Design of Experiment
- Implement
  - Code change & Testing
  - Run Experiment & Monitor
- Measurement
  - Result Measurement
  - Data Analysis
  - Decision Making





Process of A/B Test

- Design
  - Understand Problem & Objectives
  - Come Up with Hypothesis
  - Design of Experiment
- Implement
  - Code change & Testing
  - Run Experiment & Monitor
- Measurement
  - Result Measurement
  - Data Analysis
  - Decision Making

Design of Experiment

Design of Experiment (DOE)

Outline

- Key Assumptions
- Assignment
- Metrics
- Exposure & Duration
- Sample Size Calculation

DOE - Key Assumptions

- The factor to test is the only reason for difference
- All other factors are comparable
- A unit been assigned to A or B is random
- Each experiment unit are independent

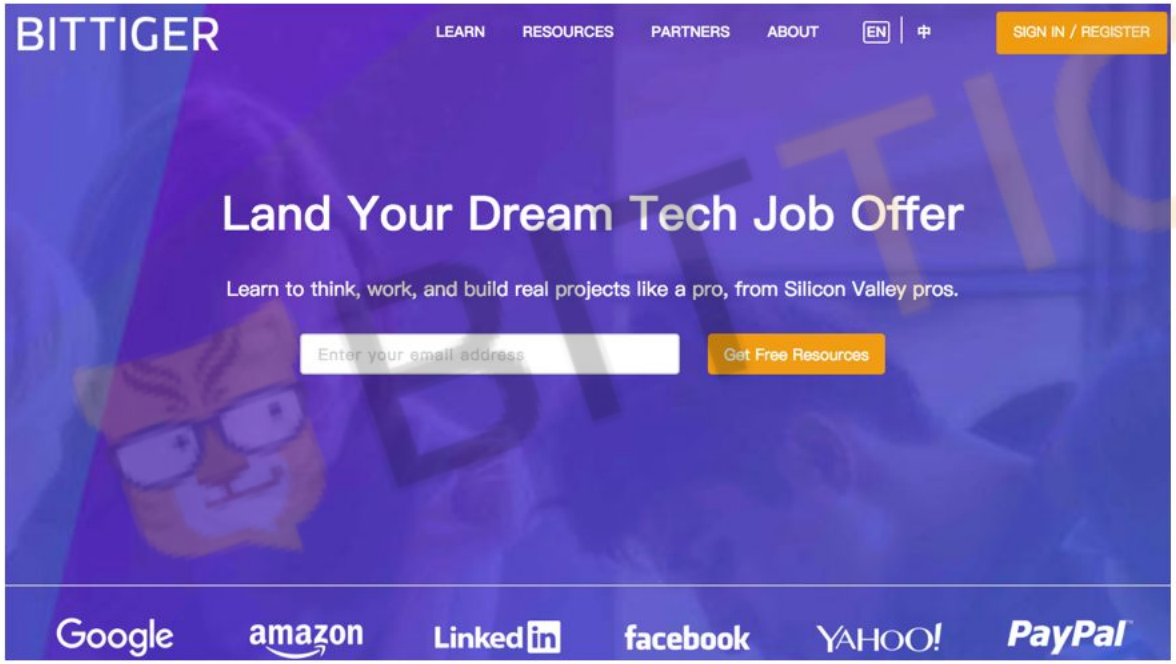
Principles of Experiment Design

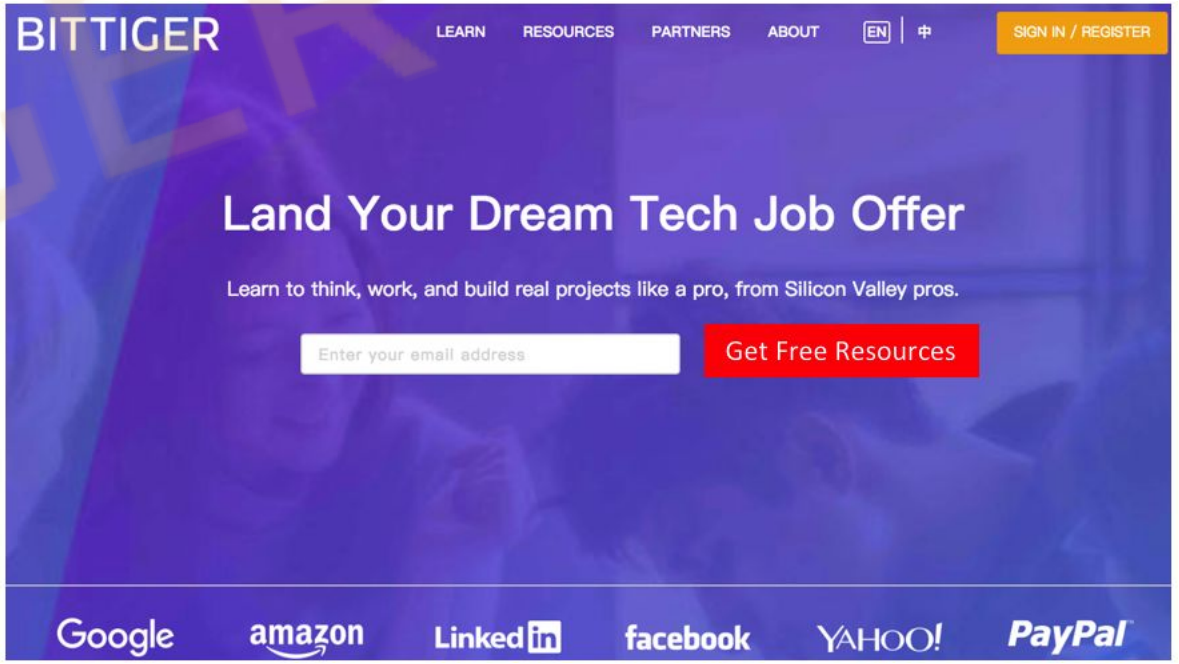
- **Independent** samples
- **Block** what you can control
- **Randomize** what you can not control

DOE - Assignment Unit

BitTiger
https://www.bittiger.io/

How to decide which version to display to whom?

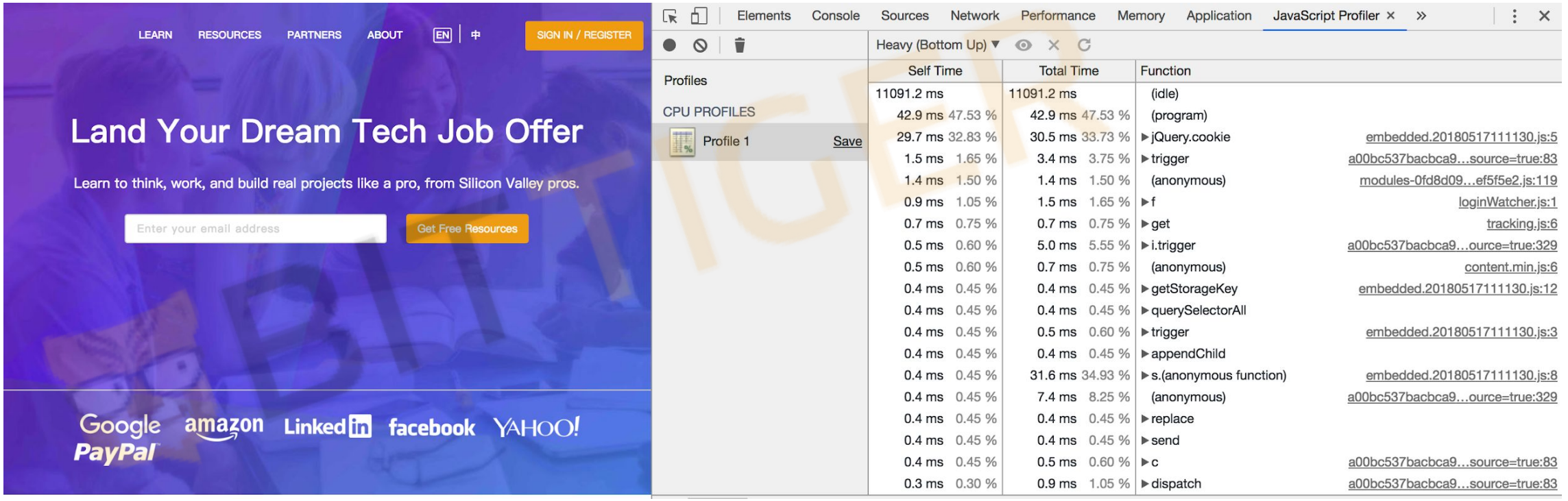




DOE - Assignment Unit

What is the unit to split A/B?

User_id? Cookie_id? Device_id? Session_id?IP address? etc



DOE - Assignment Unit

Considerations

- What are the eligible subjects we try to influence
  - Example 1: Pop up promotion ‘15% off if register today’
  - Example 2: Send emails with ‘new dress arrivals’
- What is the objective
  - Example 1: Remove homepage animation to reduce loading time
  - Example 2: Change button color to have more users to click
  - Example 3: Test impact of a change in ETA calculation method on trip cancellation rate. Rider_id, driver_id, trip_id?
  - Example 4: Test impact of a change in ETA calculation method on user retention rate. Rider_id, driver_id, trip_id?
- Independence & User experience
  - Example 1: Change homepage design in an app
  - Example 2: Add new video chat filters

DOE - Assignment Unit

In Practice, assignment unit

- Default is user_id
- Sometimes there is not only one right answer. Have to make a decision but aware of the pros & cons

Split % - % of Users in test / control

- Most common 50/50 split
- Sometimes not
  - Time sensitive e.g. holiday marketing campaign

DOE – A/A Test

- Randomly assigned
- Test / Control % is as designed
- One unit only in one group
- All other factors are comparable

How to Check?

A/A Test: use A/B test framework to test two identical versions against each other. There should be no difference between the two groups.

The goal:

- Make sure the framework been used is correct
- Data exploration & parameter estimation (e.g. sample variance)

DOE – Assignment Common Problems

Assumptions	Practical Problems	Potential Solution
Independence	Non login User	Assign by device_id, cookie_id, etc, Predict user with models
	Multi device user	
	Multiple user share a device	
Assignment to T/C is random	Bug resulting in deterministic assignment	Assignment check by group, fix bug if identified
Reproducible		Set Salt
50 / 50 split (or other % as set)	Imbalance assignment due to experiment setup	Understand why, change assignment method
Test / Control are comparable in all dimensions except treatment factor	Pre-bias (run A/A test to check)	set a different salt, Post experiment adjustment

DOE - A/A Test - Interview Quiz

Your colleague gave you two dataset and told you they are test/control groups from an A/B test. What will you do to make sure the datasets are appropriate to use?

What will you do if you find users been assigned to both test and control group? Do you have any concern?

DOE - Metrics

What to compare?

How would you know the impact of your experiment

How would you make a business/product decision with your experiment?

We need metrics!

DOE - Metrics

Metrics should be set before experiment start

- Understand what kind of changes your experiment would cause
  - Usually multiple changes happen the same time
  - Trade-offs
- Understand what are the metrics worth monitoring
- Understand the importance of these metrics
- Set expectation how these metrics would change

DOE - Metrics

What are the potential positive & negative impact of following experiments?

- Example 1: Remove homepage animation to reduce loading time
- Example 2: Change button color to have more users to click
- Example 3: Use real time traffic data to make ETA calculation more accurate to increase matching efficiency
- Example 4: Add new video chat filters to increase user engagement

DOE - Metrics

Set key evaluation metrics

- That is what you use to make a decision
- Usually one or a few
- Sometimes use a comprehensive evaluation metrics (e.g. weighted average of three metrics)

Other metrics worth monitoring

- Avoid unwanted negative impact
- Understand change in other metrics

DOE - Metrics - Interview Quiz

What metrics are you going to evaluate the success of digital ads? How to convince clients to buy ads on our website?

DOE – Exposure & Duration

Should you show the A/B version to all users?

No. may cause bad user experience if test version is bad

Start with a small proportion, like 5%, gradually roll out to more users

How long are you going to run your experiment?

In practice, we want to minimize the exposure and duration of an A/B test, because

- Optimize business performance as much as possible
- Potential negative user experience
- Inconsistent user experience
- Expensive to maintain multiple versions

DOE – Exposure & Duration

How to decide exposure %?

- Size of eligible population
- Potential impact
  - User experience
  - Business impact
  - Easy to test & debug?

  - Example 1: Redesign the layout of your app. May significantly change user behavior. Needs three teams of engineers to coordinate
  - Example 2: Change button color to have more users to click. Need one engineer 10 mins to make a change

How to decide duration?

- Minimum sample size
- Daily volume & exposure %
- Seasonality (at least one seasonal period)

Power Analysis & Sample Size Calculation

Type I, II error, power

	Ground Truth	
Decision	H_0	H_a
Not reject H_0	Correctly (not reject null)	Type II error, β
Reject H_0	Type I error, α	Correctly (reject null), power

$$\alpha = P(reject\ H_0 \mid H_0)$$

$$\beta = P(not\ reject\ H_0 \mid H_a)$$

$$power = P(reject\ H_0 \mid H_a)$$

Data Assumptions

- What distribution assumptions are you making to your data?

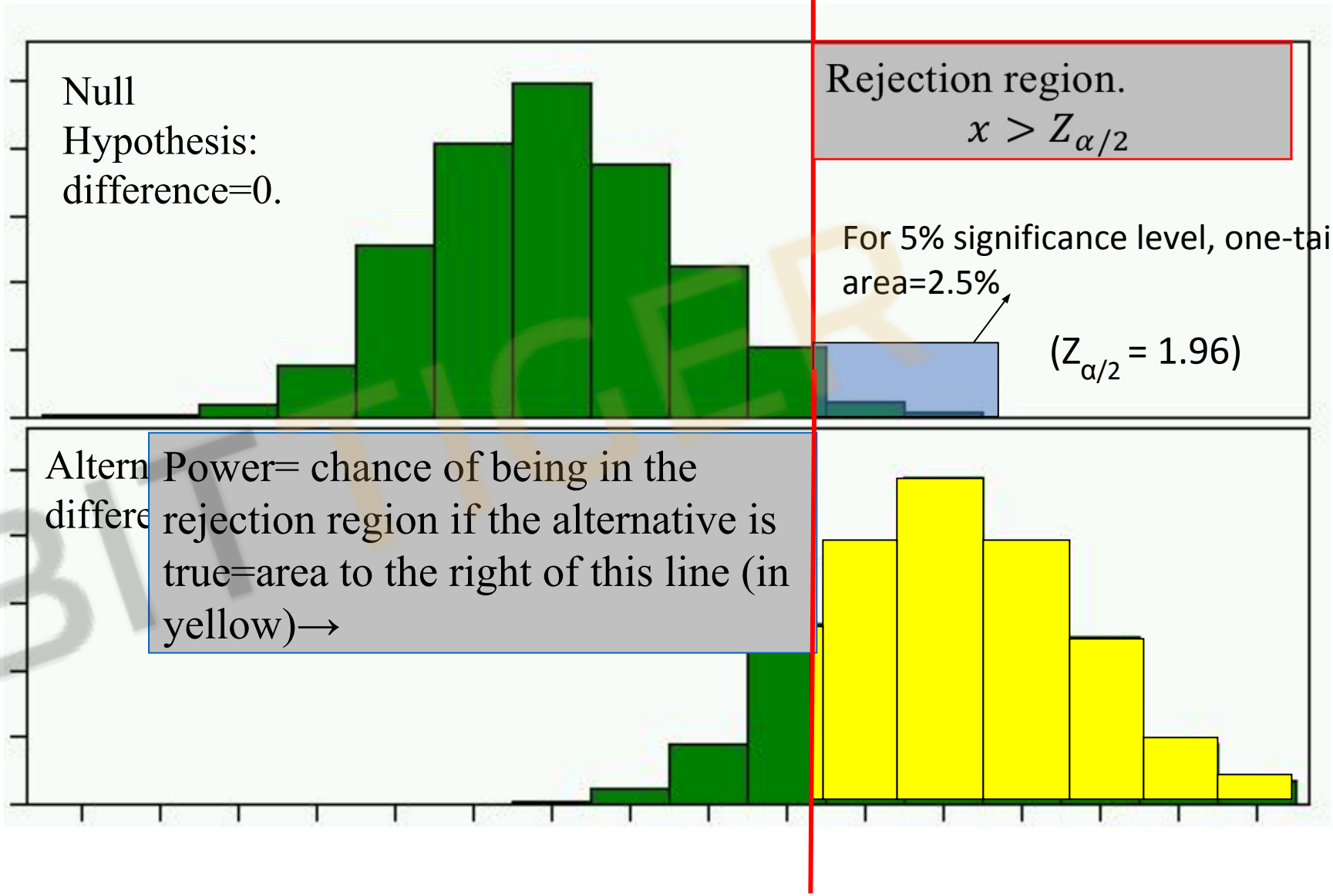
  i.i.d. Normal distribution, Central Limit Theorem

- What is the null hypothesis of your test?

$$diff = \mu_A - \mu_B = 0$$

Power Analysis

$diff \sim N(0,1)$



$diff \sim N(3,1)$

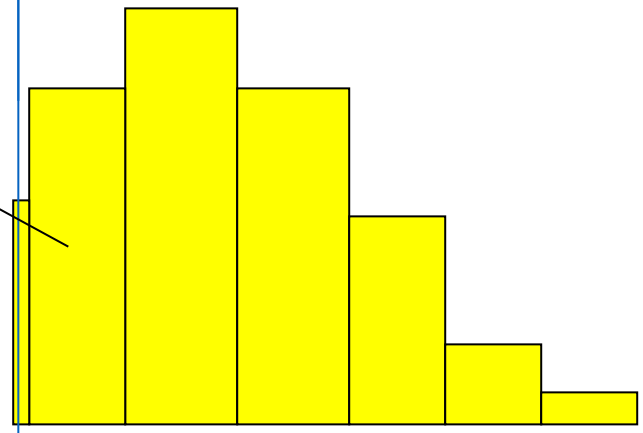
Power Analysis

Rejection region.
Any value >= $Z_{\alpha/2}$

Power= chance of being in the rejection region if the alternative is true=area to the right of this line (in yellow)

Power here:

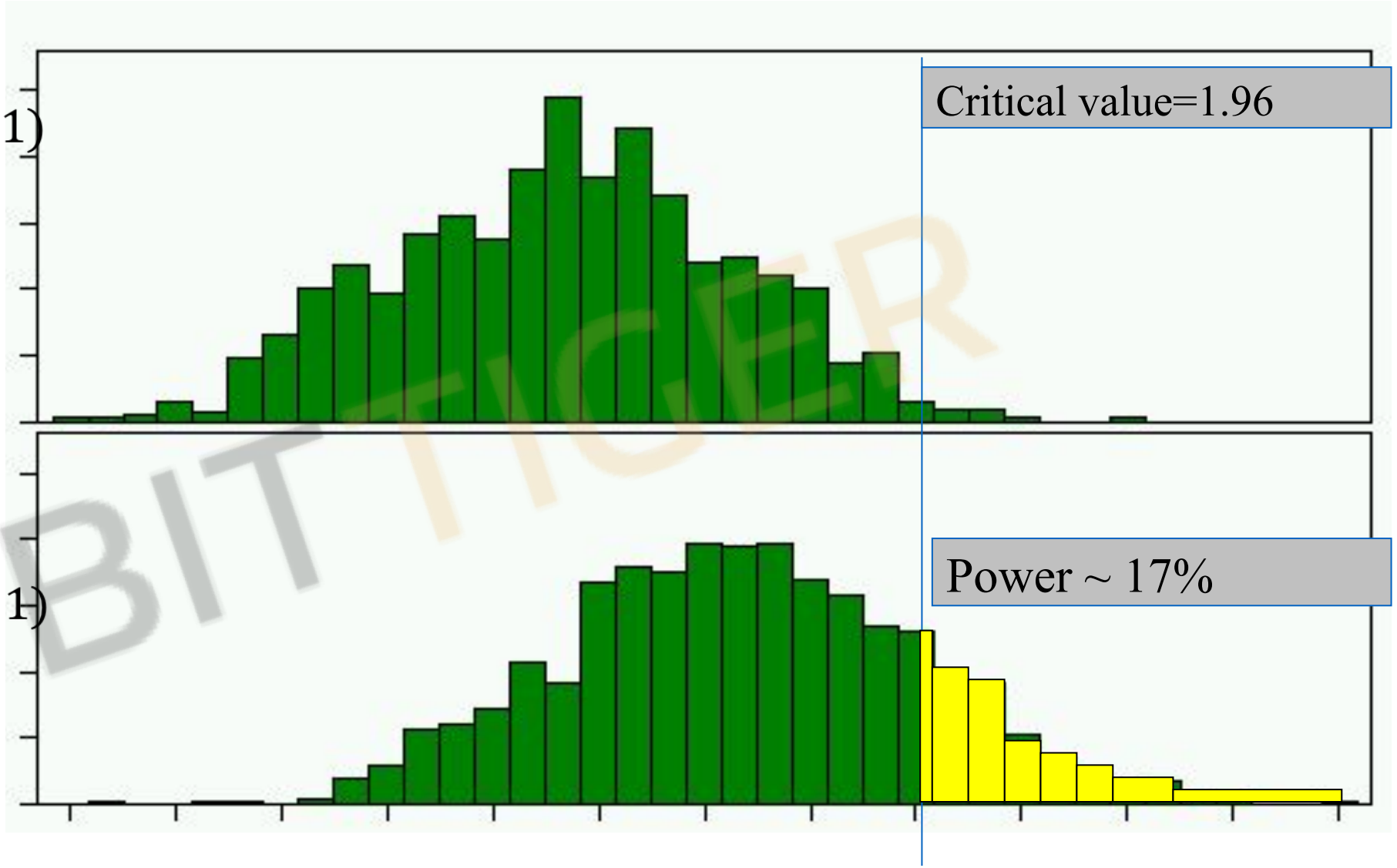
$$P(X > 1.96 \mid \mu = 3, \sigma = 1)$$
$$= P(Z > \frac{1.96 - 3}{1})$$
$$= 85\%$$



Power Analysis

$diff \sim N(0,1)$

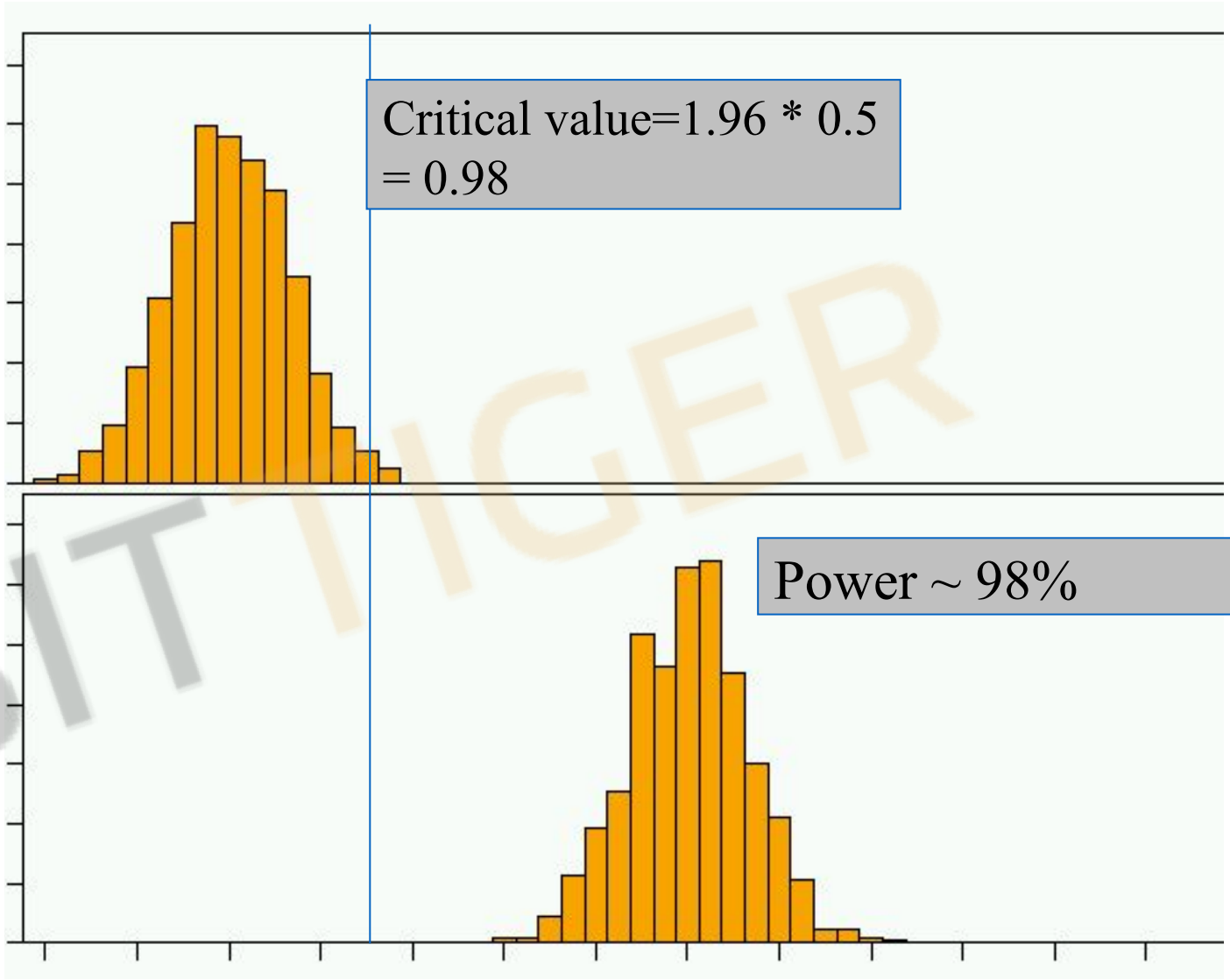
$diff \sim N(1,1)$



Power Analysis

$diff \sim N(0, 0.5)$

$diff \sim N(2, 0.5)$



Factors Impact Power

How is the power change if the following factors increase?

1. Size of the effect ↑
2. Variance of distribution ↓
3. Significance level desired α ↓

Sample Size Calculation

Represents the desired power (typically .84 for 80% power).

Sample size in each group (assumes equal sized groups)

$$n = \frac{2\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

Standard deviation of the outcome variable

Effect Size (the difference in means)

Represents the desired level of statistical significance (typically 1.96 for 95%).

Sample Size Calculation

$For\ given\ \beta\ (power), \alpha(significance\ level), \sigma\ (standard\ diviation\ of\ data)$

$$Z_\beta = \frac{critical\ value - diff}{standard\ error(diff)} = \frac{z_{1-\alpha/2} * SE(diff) - diff}{SE(diff)}$$

$$= -Z_{\alpha/2} - \frac{\text{diff}}{\text{SE(diff)}} = -Z_{\alpha/2} - \frac{\text{diff}}{\sqrt{2\sigma^2/n}}$$

$$\therefore \text{n} = \frac{2\sigma^2\left(Z_\beta + Z_{\alpha/2}\right)^2}{diff^2}$$

$$\text{SE(diff)} = \sqrt{Var(diff)} = \sqrt{\text{Var}(\overline{X_a} - \overline{X_b})}$$

$$= \sqrt{\text{Var}(\overline{X_a}) + Var(\overline{X_b})}$$ as X_a and X_b are ind

$= \sqrt{2\sigma^2/n}$ n is sample size of one group, assuming two groups have equal sample size

If not equal variance, $\text{SE(diff)} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$

DOE – Sample Size

$$n = \frac{2\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

Need to Estimate:

- Variance - estimate with sample variance
- Difference – opportunity sizing
  - Observational data
  - Qualitative result (e.g. survey, small scale test)
  - Intuition & Practical consideration (e.g. minimum effect worthing implement the change)

DOE – Sample Size - Interview Quiz

How long would you run your experiment?

What is your minimum sample size? What factors would you consider? How would these factors impact your sample size?

What will you do to balance user experience and quick learning?

What is your roll-out plan?

What will you do if your experiment take too long to run?

Implementation

DOE – Peeking

Why calculate sample size?
Can we just let the experiment run until the result is statistically significant?

No! Highly increase false positive rate

Type I error (false positive) α = 0.05 ->

When null hypothesis is true, the chance of reject H0 is 0.05

What is the chance of seeing at least one rejection having 10 tests simutaneously?

$$1 - (1 - 0.05)^{10} = 0.4$$

DOE – Monitoring

Monitor key metrics while experiment running

- Should **NOT** frequently check result
- Should **NOT** stop once result turns significant
- Wait until get minimum sample size from experiment design
- But need to monitor for alarming changes. Pause and investigate if needed

DOE – Problems & Solutions

1. What if it takes too long to get desired sample size?
    - Increase exposure
    - Reduce variance to reduce required sample size
        - Blocking – run experiment within sub-groups
        - **Propensity Score Matching**

Example:

We want to test the impact of a product change on ads' click through rate.

We know there are users who are more clicky and have higher CTR in general and users who almost never click on an ad. Is it fair to compare all users directly?

Propensity Score Matching

Procedure

1. Run a model to predict Y (CTR rate) with appropriate covariates

   Obtain propensity score: predicted y_hat
2. Check that propensity score is balanced across test and control groups
3. Match each test unit to one or more controls on propensity score:
   - Nearest neighbor matching
   - Matching with certain width
4. Run experiment on matched samples
5. Conduct post experiment analysis on matched samples

DOE – Problems & Solutions

2. What if your data is highly skewed or statistics is hard to approximate with CLT?

Example:

1. Metrics like revenue is highly skewed & have outliers
2. In risk/fraud, most transactions have no loss while some fraud transactions have very high loss

Solutions

- Transformation (hard to interpret)
- Winsorization / Capping
- **Bootstrap**

Bootstrap

Bootstrap is a resampling method. It can be used to estimate sampling distribution of any statistics, commonly used in estimating CI & p-value & statistics with complex or no close-form estimator

Procedure

1. Randomly generate a sample of size n with replacement from the original data. n is the # of observations in original data
2. Repeat step 1 many times
3. Estimate statistics with sampling statistics of the generated samples

Practice:

Use R to generate a 100 sample from Normal(3, 5).

Calculate it's theoretical & bootstrap estimate of mean & variance

Bootstrap

Pros

- No assumptions on distribution of original data
- Simple to implement
- Can be used for all kinds of statistics

Cons

Computational expensive

Interview Quiz

- Can you run an experiment and keep reading until the result is significant?
- What if your key metrics dropped by 5% on first day? What if dropped by 20%?
- What is bootstrap? Boosting?
- The metrics of interest is 90th quantile of users' spending. How to estimate sample mean and variance?

Result Measurement

Result Measurement

- Data Exploration
  - Imbalance Assignment
  - Mixed Assignment
  - Sanity Check
- Hypothesis Test
  - Conduct test
  - Multiple Testing
- Result Analysis
  - Pre-bias Adjustment
  - Analysis unit different with Assignment Unit
  - Cohort Analysis

More Advanced Topics

RM – Data Exploration

- Data Exploration
  - Check for % of test/control units. Is the % matching DOE?
    - IF not match, need to figure out what's the cause
  - Check for mixed assignment
    - It's hard to resolve. If # of mixed samples is small, OK to remove. If big, need to figure out what's the cause
    - What's the problem of throwing away mixed samples?
  - Sanity Check
    - Are test/control similar in other factors other than treatment?

RM – Hypothesis Test

- Set up the right test
  - Mostly use T-test
  - When variance is known is large, can use Z-test
  - When sample size small can use non-parametric methods
  - For complicated statistics, can use bootstrap to calculate p-value

RM – Decision Making

- If all metrics move positively
  - Meet expectations? Yes, ready to launch
  - Be cautious if result is too good. May need to investigate (e.g. outliers)
- If some metrics move negatively
  - Are they as expected? Are these metrics important?
  - Deep dive to find causes
- If results are neutral
  - Slice / Dice on sub-groups

RM – Interview Question

- What if your result show positive impact on some metrics and negative impact on some other metrics?
- What if your result is neutral?
- What if you result is statistically significant but the margin is very small?
- Take home challenge

Multiple Testing

What if you have multiple test groups?

	Image	Headline
VERSION 1		"ACME WIDGETS"
VERSION 2		"ACME WIDGETS"
VERSION 3		"THE ONE AND ONLY ACME WIDGETS"
VERSION 4		"THE ONE AND ONLY ACME WIDGETS"

False positive rate is much higher when doing multiple testings!!
Need to control family-wise false positive rate

https://www.optimizely.com/optimization-glossary/multivariate-test-vs-ab-test/#

Multiple Testing Adjustment

Bonferroni Adjustment:

Assume we have m tests, Set $\alpha_i = \frac{\alpha}{m}$ for each experiment.

This guarantee the overall FP < α, but too conservative

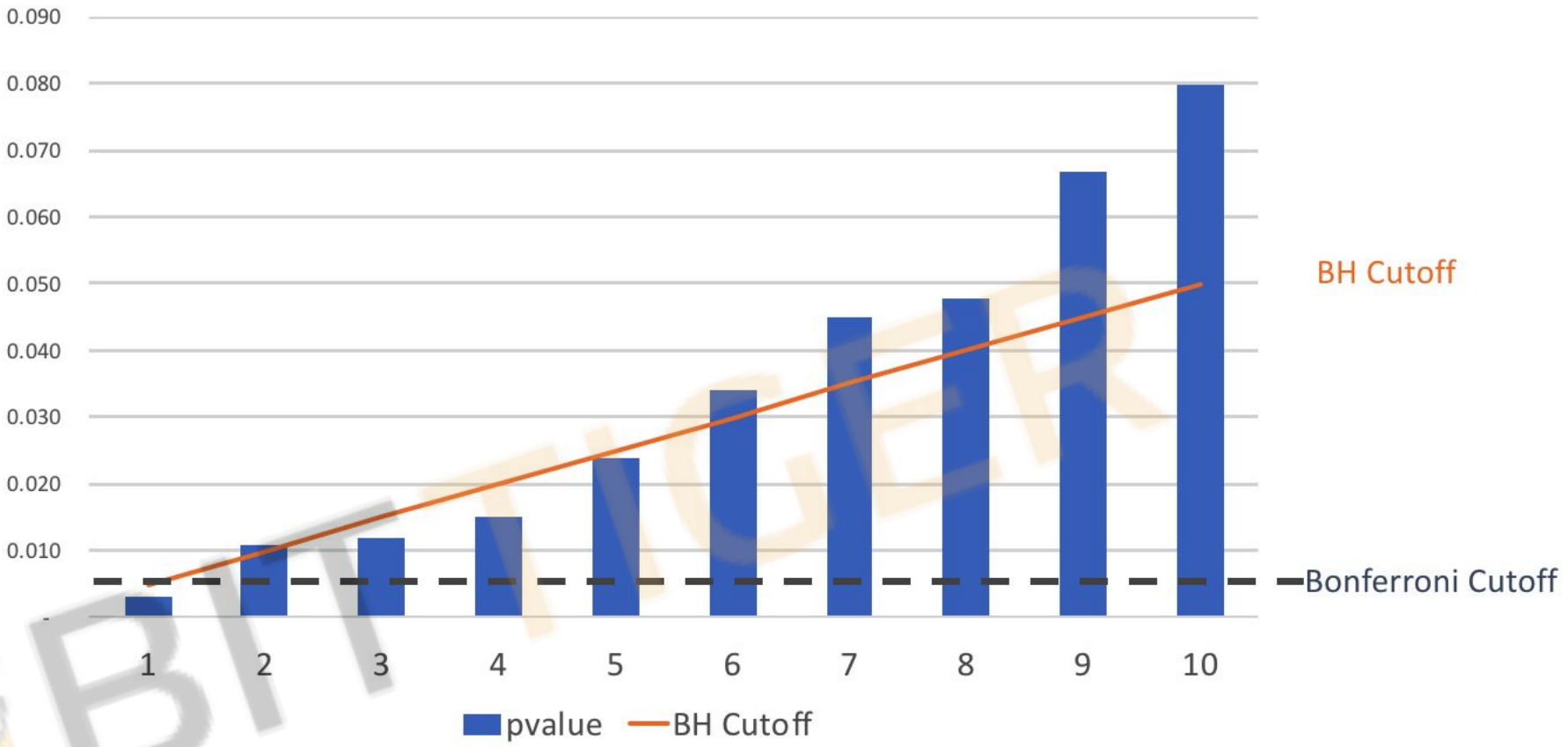
Multiple Testing Adjustment

FDR (false discovery rate) Adjustments

Benjamini – Hochberg Adjustment:

1. Rank p-values P_i of m tests from low to high
2. Find the largest k such that $p_k \leq \frac{k}{m} * \alpha$
3. Reject experiments 1,…,k. Accept experiments k+1,…, m

Multiple Testing Adjustment



Bonferroni Method: Reject Test 1, accept all rests

BH Method: Find max k that $p_k \leq \frac{k}{m} * \alpha$, for this case, k = 5. Reject T1 to T5, accept T6 to T10

Multiple Testing Interview

- 第一个人 组里比较 senior的感觉 给我介绍了一个他们之前做的问题大概就是Neyman Rubin 模型 但是记录的output有很多 要判断这些output中有没有任何一个的均值是明显有差异的 也就是一个multiple testing的问题 用 Bonferroni correction。然后问我实际发现correct之后没有一个p-value比threshold小 但是很多很接近 那么有没有办法来处理。
- 第二个人 介绍了另外一个case 就是在做ab testing的时候有可能会treatment组 出错 所以希望尽量避免这种情况 但是另一方面 又希望可以把尽量多的人放进去 如何formulate这个trade off （type1 type2 error）

Pre-bias Adjustment

We had the assumption that the A/B groups have no difference before experiment. What if the does exist difference?

Regression Adjustment

Y_post = beta_pre * Y_pre + beta_t* Treatment_Group

Diff - in – Diff Comparison

(Y_post/t – Y_pre/t) – (Y_post/c – Y_pre/c)

Cohort Analysis

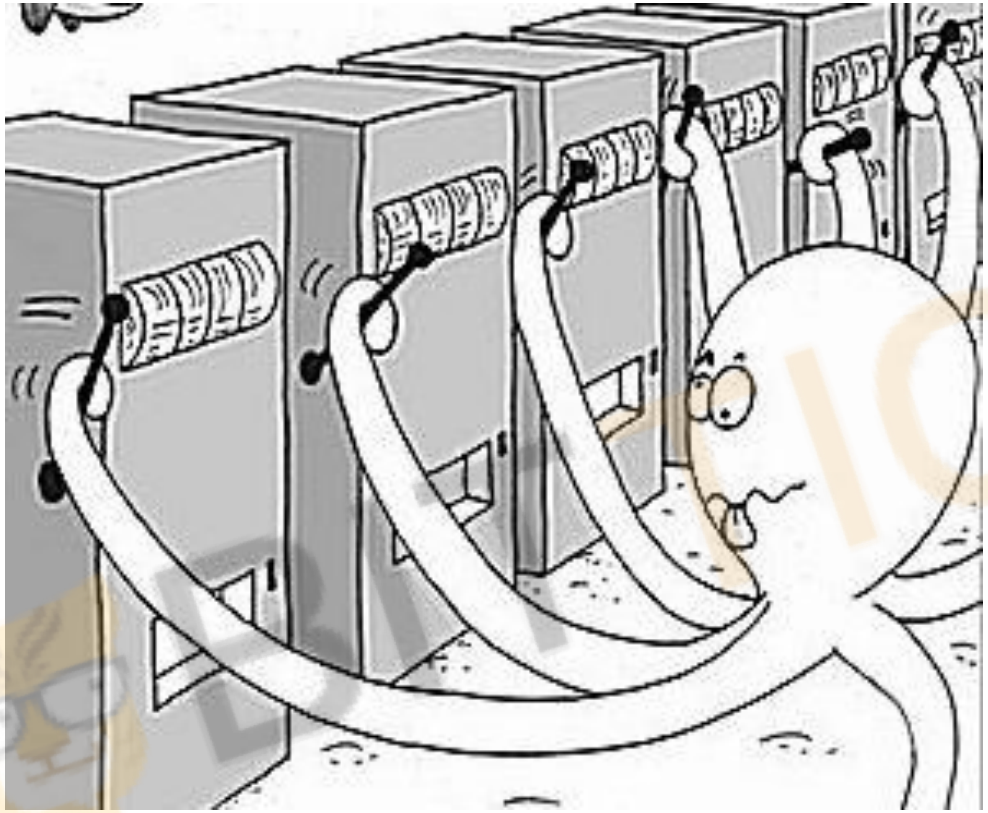
How to measure impact over time?

E.x. Spotify is testing a new recommendation system algorithm, which is expected to give more accurate recommendations thus improve user engagement. You do not expect users to notice the difference and take any actions since day 1. But users are expected to gradually get more engaged over time

Cohort Analysis

Select a cohort of users (e.g. T/C users assigned on the same day) and monitor their metrics change over time

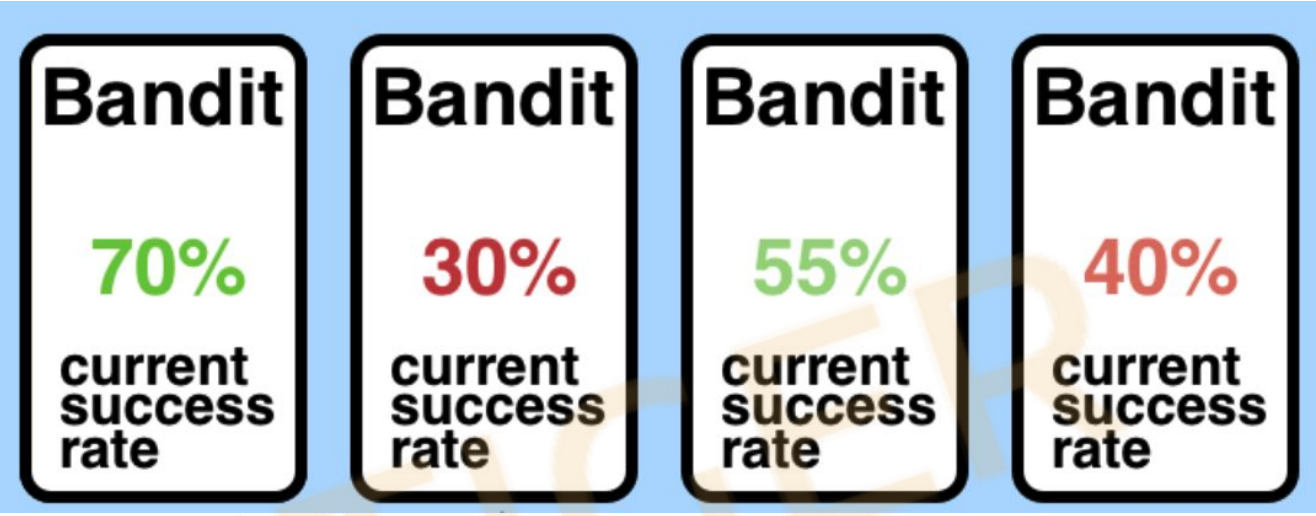
Multi-arm Bandit Problem



- Each slot machine has different rewards
- Objective: to maximize rewards in casino

Which arm would you pull?

Multi-arm Bandit Problem

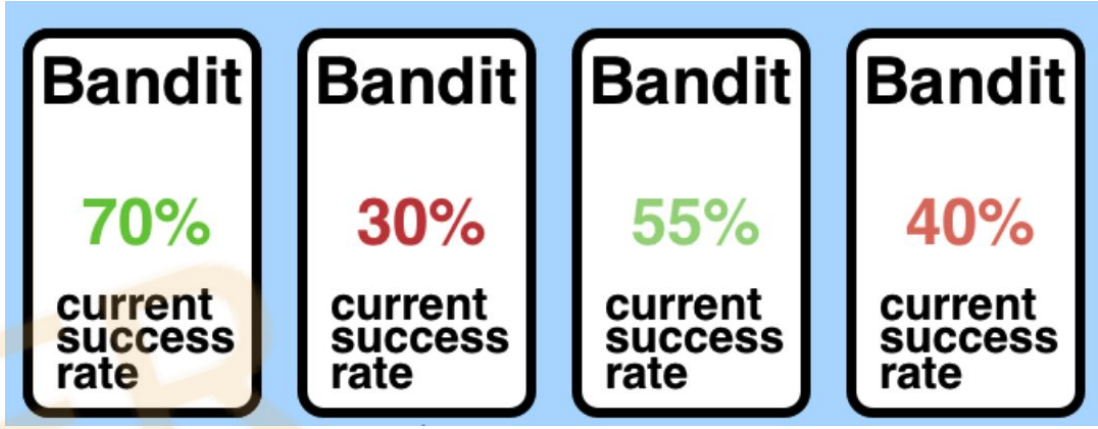


- You have multiple test groups. Each group has different success rate and are unknown before your experiment. Your goal is to maximize the overall success rate of all users. How would you allocate your users?

Multi-arm Bandit Problem

Two stage

- **Exploration**: Example Split 10% of all of your users equally between all of the treatments
- **Exploitation**: Example Use the other 90% of the tokens in the machine that rewards you the most



Different Methods:

Epsilon-Greedy: the rates of exploration and exploitation are fixed

Upper Confidence Bound: the rates of exploration and exploitation are dynamically updated with respect to the rewards of each arm

Thompson sampling: the rates of exploration and exploitation are dynamically updated with respect to the entire probability distribution of each arm

https://www.searchenginepeople.com/blog/16072-multi-armed-bandits-ab-testing-makes-money.html

Limitations of A/B Test

- Highly rely on your hypothesis
- Good for optimize small changes, Not good for innovative changes, long term strategies
- Other factors involved: e.g. learning effect, network effect

Other ways to make product improvement

- Qualitative Studies – Survey, focus group
- Observational studies

Typical Interview Examples

A/B Test in Interviews

How are A/B test questions asked?

- Asked directly
- Asked in case studies / product sense questions (Most common)
- Asked in take home projects

Example Question 1

Survey showed teenagers are less engaged with Facebook after their parents join FB. What to do?

1. Understand problem, define population & objectives & metrics
2. Brainstorm features to consider
3. How do you know if your feature works?

- Design of experiment (metrics, assignment)
- Duration & exposure of your experiment
- How would you make decision?
- What if you see xyz?

Example Question 3

You have 1M budget to spend on holiday campaigns. Possible investments include mailed ads / emails / display ads / search engine / social media ads. How would you optimize the budget?

1. Define objectives & metrics
2. Design of experiment
3. Result measurement
4. Slice / dice analysis
5. Multi-armed bandit

Example Question 3

An engineer suggest promoting new sellers on your website to boost seller growth. But another engineer is worried about it hurting overall sales. How would you make a decision?

1. Define objectives & metrics
2. Target metrics vs monitor metrics
3. Result measurement
4. Decision making with mixed results

Example Question from students - 1

1.slow roll out: How do you explain the result

Table 1: Conversion Rate for two days.
Each day has 1M customers, and the Treatment (T) is better than Control (C) on each day, yet worse overall

	Friday C/T split: 99%/1%	Saturday C/T split: 50%/50%	Total
C	$\frac{20,000}{990,000} = 2.02\%$	$\frac{5,000}{500,000} = 1.00\%$	$\frac{25,000}{1,490,000} = 1.68\%$
T	$\frac{230}{10,000} = 2.30\%$	$\frac{6,000}{500,000} = 1.20\%$	$\frac{6,230}{510,000} = 1.20\%$

Example Question from students - 2

2.Interpret AB testing result, Treatment effect for each group as below:

- Trt1: -5, CI: (-7.5, -2.5)
- Trt2: -15, CI: (-17, -13)
- Trt3: -12, CI: (-28, -4)

How to interpret C.I.? Which treatment to choose? Increase test power / accuracy?