1. What is A/B Tests
   1. An A/B test is an experiment:
      1. All elements are constant
      2. Except for one variable
      3. Variants:
         1. Different versions of a product
         2. 2 variants:
            1. A/B Tests
         3. More than 2 variants
            1. A/B/N Tests
   2. Goals:
      1. Make data driven decisions when results are reliable and repeatable
      2. Make result reproducible:
         1. Change to product 🡪 change to metrics
         2. Example
            1. Change to color 🡪 change to engagement assuming others stay the same
   3. Steps:
      1. Prerequisites
         1. Define key metrics
            1. Overall Evaluation Criterion (OEC)
            2. Example

Color of check out button affects revenue

OEC: revenue/user/month

* + - 1. Changes need to be easy to make
         1. Difficult changes introduce complexity
         2. Example: Re-designing the entire website may be too difficult
      2. Have enough “Randomization Units”
         1. Randomization Units

Who or what we randomly assign to each variant of A/B test

* + - * 1. How many is enough?

Thousands

Larger the units 🡪 Smaller changes

* + 1. Experiment Design
       1. What population to select
          1. Specific population VS. All users
          2. Example:

Users only available in a geographic region

* + - 1. Size of an experiment
         1. Statistical power 🡪 sample size
      2. How long to run an experiment
         1. Seasonality
         2. Day of week
         3. Primacy and novelty effect
    1. Running Experiment
       1. Collect Data
          1. Instrument loggings
          2. Utilize companies’ platform
    2. Result to Decision
       1. Check & interpret results to make decisions
       2. Sanity checks
          1. Passed 🡪 continue with analysis
          2. Failed 🡪 discard the results, look into root cause
       3. Consider
          1. Tradeoffs between different metrics

Example:

User engagement GOES UP

Revenue GOES DOWN

* + - * 1. Cost of launching a change

Example:

Engineering maintenance

Opportunity cost

* + - * 1. When costs are high:

Benefit should outweigh the cost

Set practical significance boundary

* + - * 1. When costs are low:

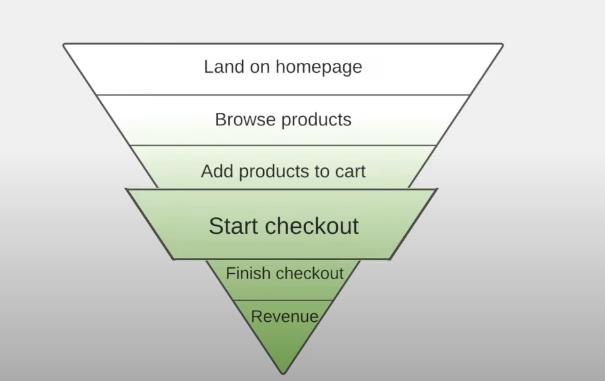
Launch any positive changes

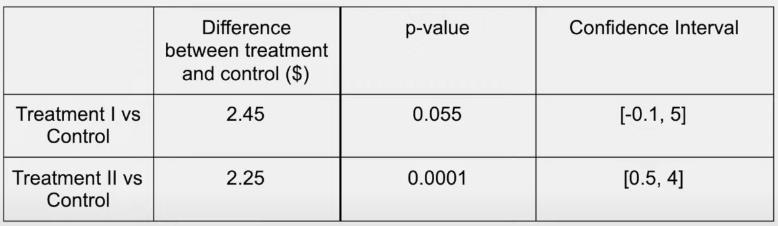
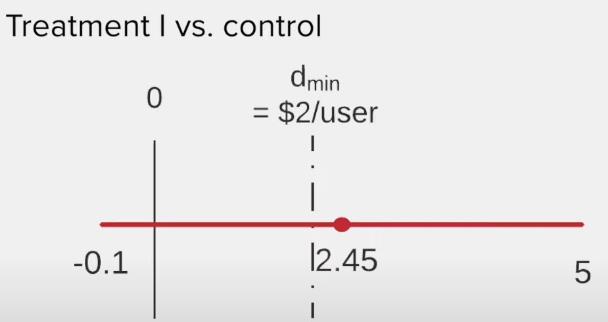
* + - 1. Post-launch Monitoring
         1. Long-term effect

Different from short-term effect

Example:

Novelty effect

1. Real-life Example
   1. Prerequisites
      1. Objective & Key Metrics
         1. Key Metric
            1. Revenue
            2. Revenue per User
         2. Product Variants
            1. Control Group: Stay the same
            2. Treatment Group1: Similar Products on the checkout page
            3. Treatment Group1: Similar Products on popup windows when checkout
         3. Randomization Units
            1. Users
            2. Assume enough users
   2. Experiment Design
      1. Users to target:
         1. All users?
         2. Specific segment of users?
         3. 
      2. Practical Significance Boundary
         1. Assume practical Significance: Revenue increase: 2$/user
         2. Power of the test: 80%
         3. Significance Level: 5%
      3. Sample size =
         1. : sd of population
         2. : Difference between treatment & control
         3. Assume sd = 20, then sample size = 1600 and 4800 unique users for 3 variants
         4. Smaller change means more samples
         5. Smaller significance level like 2.5% also means more samples
      4. Ramp-up plan:
         1. No bugs
         2. Traffics can be handled
         3. Expose to a small population
         4. Gradually increase percentage
         5. Chart, bar chart

            Description automatically generated
      5. Day of week effect
         1. People behaves differently
         2. Run experiment for >= 1 whole week
      6. Seasonality
         1. Data during holiday cannot be used for analysis
      7. Primacy and novelty effects
         1. Users respond to changes differently
   3. Result to Decision
      1. Sanity Check
         1. Number of users assigned to groups is truly random
         2. Latency when loading the webpage to make sure there is no significant latency difference between these groups
      2. Hypothesis Testing
         1. Recommend launching a change when
            1. Statistically significant
            2. Practically Significant
         2. Example:
            1. 
            2. 
            3. Confidence Interval overlaps with 0: not statistically significant
            4. Point estimate > practical significance boundary: practically significant
            5. In this case, run a follow-up test with more power.
2. Hypothesis Testing in A/B Testing
   1. Experiment: test of color of a button
      1. Click through probability: N(users who clicked) / N(total users)
      2. 1000 users in both control & treatment groups
   2. Result
      1. Control group: 1.1% CTP
      2. Treatment group: 2.3% CTP
   3. Steps for Z-distribution
      1. Which hypothesis test to use?
         1. Bernoulli population: either clicks or doesn’t click
         2. Control group: n \* = 1000 \* 1.1% = 11
         3. Treatment Group: n \* = 1000 \* 2.3% = 23
         4. Test statistic follows Z-distribution
         5. Measurements
            1. Users clicked ,
            2. Total number of users ,
      2. What is the null hypothesis?
         * 1. We would expect
           2. Test statistics:

Estimate Standard Error

Choose a SE can represent both groups

“Pooled” Standard error

“Pooled” probability of click,

Total Probability across 2 groups

= 0.00578

* + 1. Is the result statistically significant?
       1. Critical z-score (α:0.05) = 1.96
       2. TS > 1.96 or TS < -1.96, reject the null hypothesis
       3. In this example:
          1. TS = 2.076 > 1.96
          2. Test is statistically significant
    2. Is the result practically significant?
       1. Confidence interval of d
          1. Center of C.I. = 0.012
          2. Width of C.I. (margin of error)

M = Z \* = 1.96 \* 0.0578 = 0.0113

CI of d: 0.012 +- 0.0113 = 0.0007 ~ 0.0233Chart

Description automatically generated

In this case, the point estimate is greater than the significance boundary, but the left end of CI is less than significance boundary.

Best guess: there is a practical significant change, but it is possible the change is not practically significant

* + 1. Make decisions
       1. Not confident the change is practically significant
       2. Not recommend launching the feature
    2. Few Points
       1. Check if CI overlaps with 0
          1. If it does, result is not statistically significant
       2. Equivalent to comparing TS with critical value
  1. Two Sample T-test
     1. Calculate means for control and treatment group
     2. Determine Practical Significant boundary and Significance Level α
     3. Variances
        1. 2 groups have similar variances
           1. Compute “Pooled” Variance
        2. 2 groups have different
           1. Compute “unpooled” Variance
     4. Null hypothesis
        1. Test statistic with “pooled” variance
        2. “Pooled” standard error
     5. Statistically Significant
        1. Calculate TS
        2. Critical t-score value (α = 0.05, df =):
        3. Compared TS to t-score
     6. Practically Significant
        1. Calculate CI
        2. Compare CI with practical significance boundary
     7. Make Decision
  2. Welch’s T-Test
     1. Used when
        1. Two samples have different variance and/or sample sizes
        2. Compute “Unpooled” standard error
        3. Confidence Interval of d
           1. A picture containing text

              Description automatically generated

1. Sample Size Estimation
   1. Rule of thumb
      1. : Variance
      2. : Significance Level, Type I error (False Positive)
      3. : Type II error (False Negative) = 1- power
      4. : Difference between two groups
   2. Alpha (Significance Level)
      1. Smaller alpha 🡪 Higher confidence Level 🡪 More sample needed
      2. As our sample size increases our uncertainty decreases and we have a greater confidence in our estimation.
   3. Beta (1 – Power)
      1. Usually set as 0.2
      2. Smaller Beta 🡪 Greater Power 🡪 More sample needed
   4. Variance Estimation
      1. : Estimate of Variance
      2. From historical data and tests:
         1. Query from historical logs
         2. Previous A/B tests and A/A tests
         3. Run an A/A test when no historical data is available
   5. Delta
      1. : Difference between Control and Treatment
      2. Use minimum detectable effect (A.K.A practical significance)
         1. Example: 10M increase in revenue, 10k increase in button click
         2. Smallest change that is meaningful to the business
      3. Need more data to detect smaller changes
         1. Large sample sizes increase the probability of estimating the metrics accurately
2. Choosing the right Metrics for A/B Testing
   1. Goal Metrics
      1. Company’s long-term vision& mission
      2. Reflect what the company cares about and simple to communicate with stakeholders
      3. Stable over time
      4. May not be suitable for experiments:
         1. Difficult to measure
         2. Not sensitive to product changes
         3. Example:
            1. Facebook cares about revenue but not every team uses it for experiments
            2. Teams may work on improving user engagement or App & web performance
   2. Driver Metric
      1. Surrogate metric/ indirect/ predictive metric
      2. Align with goal metric but more sensitive and actionable (short-term)
      3. Example:
         1. Marketing team Goal: acquire new users
         2. Driver metric: # of new users/day
   3. Guardrail Metric
      1. Organizational guardrail metric
         1. Negative 🡪 business suffers loss
         2. Example:
            1. Loading time increase a few ms 🡪 loss of customers & revenue
            2. Metrics example:

page loading latency

Errors per page

Client crashes

* + 1. Trustworthy-related metrics
       1. Monitor trustworthiness of experiments
       2. Check violation of assumption
       3. Example: Randomization units assigned to variants
          1. Numbers are different 🡪 sample ratio mismatch
          2. T-test or chi-squared test shall be performed to check the assignment ratio matches with what was designed
  1. Application of Metrics
     1. Contest is important for metrics
        1. Same metric can be used differently
        2. Example:
           1. Front-end team:

Goal: reducing latency

Driver metric: time to interactive (TTI)

* + - * 1. Product team:

Guardrail metric: time to interactive (TTI)

* 1. Attribute of good metrics
     1. Simple (usually can describe in one sentence)
     2. Clear (No ambiguity)
     3. Actionable (offer insights how to improve)
     4. Example:
        1. Short-term revenue
           1. Increase price 🡪 increase revenue
           2. Losses customers
  2. 3 attributes of metrics for experimentation
     1. Measurable:
        1. Be able to calculate metrics with data collected during the experiment period
     2. Attributable:
        1. Be able to attribute metrics to the experiment variants
     3. Sensitive and timely:
        1. Experiment metrics should be sensitive enough to detect changes in timely fashion
     4. In online experiments:
        1. A few driver metrics
        2. A few guardrail metrics
     5. How to make a decision when one metrics goes up and one metric goes down?
        1. Mental model of the trade-offs
           1. Example:

User acquisition & revenue

Expensive campaigns (discount or coupon) when doing users acquisition may degrade revenue

Need to discuss with other teams

* 1. 2 suggestions on formulating metrics
     1. Combine a few driver metrics into Overall Evaluation Criterion (OEC)
        1. Weighted combination
        2. No more than 5 driver metrics
           1. Too many metrics:

Confusion & ignoring key metrics

Increase chances of false discovery

* 1. Questions related to choosing driver metrics
     1. Sensitive and timely
        1. Good example: CTR
           1. Immediately reflects Ads performance
        2. Bad example: DAU (Daily Active User)
           1. Takes time for users to purchase, adopt to the product
           2. More like a goal metric
           3. Important to the business but not suitable for A/B tests
     2. Measurable
        1. Good example: CTR
           1. Counts can be obtained in real-time
        2. Bad example: MAU (Monthly Active User), user retention
           1. Out of the experiment timeframe
     3. Attributable
        1. Good example: CTR
           1. Attribute CTR to the Ads design
        2. Bad example: DAU, user retention
           1. Many other things can cause the change
     4. How to select metrics?
        1. Combine qualitative and quantitative methods:
           1. Qualitative: User experience research, focus groups, and surveys to understand users’ needs
           2. Quantitative: data analysis, such as analysis of logs to see what users do and find patterns in the logs
        2. Fully understand the goal of a test
           1. Be as specific as possible to fully understand the goal
           2. Example:

Is the goal about user growth?

Is it to improve engagement?

Is it to increase revenue?

Is the change about acquisition, activation, retention, referral, or revenue?

* + - * 1. Example questions: YouTube hides dislike counts on videos

Goal: protect creators, especially small creators, from harassment and make them feel safe.

What is expected: small creators become more engaged

Two driver metrics:

Average time spent on YouTube per creator

Average number of videos published per creator

* + - 1. Analyze user experience
         1. Consider the steps users in each group need to take to use feature or a product
         2. Most products or features have a funnel that moves users towards taking key actions or desired outcomes that are meaningful to the business
         3. Example questions: YouTube hides dislike counts on videos

Desired outcome: fewer dislike on videos on smaller channels

Control: Viewers can see the number of dislikes

Treatment: Viewers cannot see dislike numbers, but feedback is possible

Metric: Average number of dislikes per viewer

Measure the average number of dislikes for smaller creators (ideally decreased)

What YouTube found:

Reduction in dislike attaching behavior

1. Choosing Randomization Units for A/B Tests
   1. Randomization Units (Unit of Diversion)
      1. Who or what we randomly assign to each variant of A/B test
      2. Impacts user experience and metrics
      3. Not only users can be used
      4. Commonly use
         1. User ID
            1. Pros:

Stable across time and platform

* + - * 1. Cons:

Mindful of confidentiality

Requires to login to identify

* + - 1. Cookie
         1. ID specific to browser and device
         2. Small footprints that can expire
         3. Used to identify users without logging on
      2. Event
         1. E.g. page view and session
         2. A finer level of granularity than a user ID
         3. One user can be connected to many page views or sessions
         4. Page-level randomization

Every page visit is a randomization unit

Do not require users to log on

Don’t distinguish between users

* + - * 1. Session-level randomization

Continuous period of activities

Expires after 30-min inactivity

Treats every session as independent occurrence

One user can be assigned to different variants

* + - * 1. Typically, more page views and sessions than users

Provides more units and gives us more power to detect small changes

May lead to an inconsistent user experience

Don’t use event as randomization unit if the change is visible to users

* + - 1. Device ID
         1. Immutable ID associated with a specific device
         2. Only available for mobile devices
  1. How to choose randomization units in different scenarios
     1. Consistent user experience
        1. Major change of UI or workflow 🡪 User-level randomization REQUIRED
        2. Non-user visible changes (website performance changes, changes in back-end algorithms) 🡪 User-level randomization NOT REQUIRED
     2. The coarseness of the randomization unit and unit of analysis
        1. Recommendation: Randomization unit is at least as coarse as unit of analysis (Metric)
        2. E.g.
           1. Metric: Click-through rate
           2. Unit of analysis: page view
           3. Randomization unit: User
           4. Randomization unit is coarser than unit of analysis
        3. E.g.
           1. Metric: Avg number of clicks per user
           2. Unit of analysis: User
           3. Randomization unit: Page view
           4. Randomization unit is less coarse than unit of analysis
           5. Makes computing user-level metrics unmeaningful

1. Common questions in DS interviews
   1. What is A/B testing?
      1. A/B test (Controlled Experiment)
         1. Used in industry to make decisions
         2. Simplest from: control A, treatment B
            1. Control: existing features
            2. Treatment group: new features
      2. Evaluate features with a subset of users
   2. Designing an A/B test
      1. How long to run an A/B test?
         1. Determine the sample size
            1. Type II error or power
            2. Significance level
            3. Minimum detectable effect
            4. Simple Formula:

is obtained from data

How to estimate

We don’t know before experiments

Use minimum detectable effect

* + - 1. Use sample size and number of users
      2. Round the duration by weeks to capture the weekly pattern
  1. Multiple Testing problem
     1. Test multiple variants of a feature
        1. Colors
        2. Homepage design
        3. Sample question:
           1. 10 tests are running with different landing page
           2. 1 case won and the p-value < 0.05
           3. Should you make the change

The answer is NO

Should not use the same significance level

In the situation od more than 2 variants, probability of false discovery increases

Example: there are 3 groups, what is the change of at least one false positive?

P(no false positive) = (1-0.05)^3 = 0.95^3 = 0.857

P(at least 1 false positive) = 1 – P(no false positive) = 0.143

Type 1 error over 14%

* + - 1. Bonferroni Correction
         1. Significance level/ number of tests
         2. Significance level of 10 tests = 0.05/10 = 0.005
         3. The method is too conservative
      2. Control False Discovery Rate (FDR)
         1. FDR =
         2. Example: With 200 metrics with FDR at 0.05

5% False positive

At least 1 false positive in 200 metrics

* 1. Novelty and Primacy Effect
     1. Primacy effect (Change aversion):
        1. People are reluctant to change
     2. Novelty effect:
        1. People welcome the changes and use more
     3. Effects will not last long
     4. An A/B test has larger or smaller initial effect
        1. Due to novelty or primacy
     5. Sample question:
        1. Ran an A/B test on a new feature
        2. The test won and we launched the change
        3. After a week, the treatment effect quickly declined
        4. Answer:
           1. Novelty effect
           2. Repeat usage declined when effect wears off
     6. Methods to deal with those effects
        1. Rule out the possibility:
           1. Run tests only on first time users
        2. If test is already running:
           1. Compare first time users to old users in treatment group
  2. Interference Between Variants
     1. Typical Design:
        1. Split users randomly
        2. Users are independent
     2. Cases when assumption fails:
        1. Social network: Example: Facebook, LinkedIn, Twitter
        2. Two-sided markets: Example: Uber, Lyft, Airbnb
     3. Sample question:
        1. Test a new feature to increase posts created per user
        2. Assign each user randomly
        3. The test won by 1% in terms of posts
        4. What would happen after new feature is launched to all users?
           1. Will it be same as 1%?
           2. If not, would it be more or less?
           3. Assume no novelty effect
           4. The answer is the difference will be more than 1%

Network effect:

User behaviors are impacted by others

The effect can spillover the control group

The difference underestimates the treatment effect

Two-sided markets:

Example: Uber, Lyft, Airbnb

Resources are shared among control and treatment groups

Example: treatment group attracts more drivers, less drivers are available for control group

Actual effect < treatment effect

* 1. Dealing with interference
     1. Sample questions:
        1. A new feature provides coupons to our riders
        2. Goal: increase rides by decreasing price
        3. Testing strategy: evaluate the effect of the new feature
     2. Main idea:
        1. Isolate users
     3. Two-sided markets:
        1. Geo-based randomization:
           1. Split by geolocations
           2. Example: New York vs. San Francisco
           3. Big variance since markets are unique
     4. Time-based randomization:
        1. Split by day of week
        2. Assign all users to either treatment or control
        3. Only when treatment effect is in short time
        4. Works:
           1. Treatment effect lasts in short time
           2. Uber’s surge price
        5. Does not work:
           1. Treatment effect takes a long time
           2. A referral program (usually takes a long time)
     5. Social network:
        1. Create network clusters:
           1. People interact mostly within the cluster
           2. Assign clusters randomly
        2. Ego-network randomization:
           1. Originated from LinkedIn
           2. A cluster is composed of an “ego” and her “alters”
           3. One-out network effect: user either has the feature or not
           4. It’s simpler and more scalable
     6. Conclusion:
        1. All methods have limitations
        2. Evaluate method based on scenario
        3. It’s possible to combine methods