A/B testing key takeaways

Key Assumptions For A/B testing-baidu
Units are assigned randomly
Units are Independent

What is A/B testing

Definition

A/B testing (also known as split testing) is the process of comparing two versions of a web page, email, or other marketing asset and measuring the difference in performance. You do this giving one version to one group and the other version to another group.
Establish causal relationship between actions and results
Product iteration
- UI

How to ensure features tests are independent?
Only to make sure each team have same proportion to other feature test
Marketing optimization

Assumption:
- Independent samples
- Block what you can control
- Randomize what you can not control

- 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

Cons of AB testing

- quantitative metrics
Needs to dive into different groups of people
- short terms vs long terms
Only short term effects
Concern with long-term metrics
- New experiment(aversion & novelty effect)

Statistical foundations

- Normal distributions

Normal Distribution

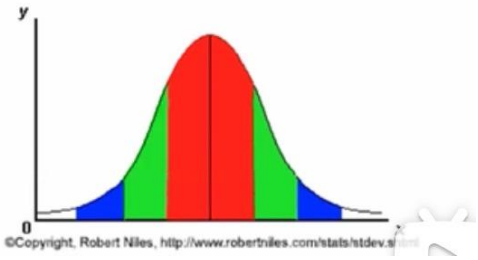
- Normal distribution

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



Proportion is binomial distribution

- Central Limit Theorem

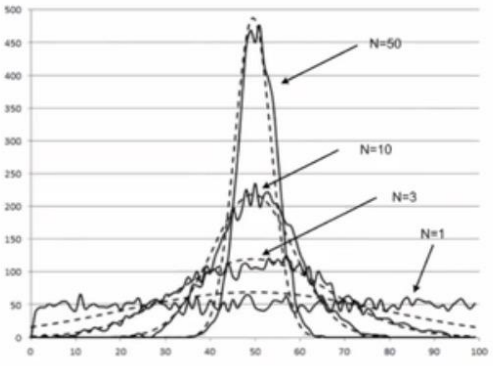
When sample size is big enough,

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

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

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

- Application
  - Binomial distribution



When doesn't work?

<u>N<30 --- T distributions</u>

<u>T-distribution</u> more conservative, heavily two tailed, population variation is unknown

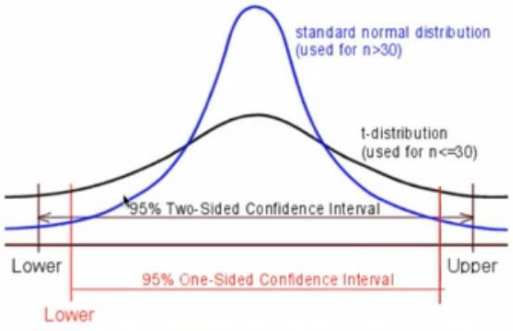
<u>Z-distribution</u> N is more larger,population variation is known, porpotion quesiton(bernouli trail is known as variance p(1-p))

- 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

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



Highly Data skewed

Why?

Metrics is highly skewed

Risk/fraud-- high loss

90% percentile

Solutions:

Lower the variability(from absolute value to dummy variable)

Transformation (hard to interpret)

Winsorization/capping(make the large values to smaller group; shrink the outliers)

Bootstrap

Re-sampling method:

Process:

1. randomly generate a sample of size n with replacement from the original data.
2. Repeat step 1many times
3. Estimate statistics with sampling statistic

Pros:

No assumption on distribution

Simple to implement

Can be used for all kinds of statistics

Cons:

Computational expensive

- Correlation != Causal inference

Observational study: Before experiment, how can you tell the X causing Y (find out the confounding variable)

Randomized experiment: randomization other variables but control one variable

Process

- Define goals and funnel
- Define metrics
- Form hypothesis
- Formulate the test plan
- Create variations
- Run experiments
- Analyze the test results
- Conclude

2. Design A/B testing?

(1) Design goal and funnel

What is the goal for this features?
What is the behaviors changed from the feature change?

Trade-off
Cost: inconsistent with other

(2) How to choose metrics

2.1 Goal Metrics

Company's vision& missions
Critical few metrics:
Robinhood- democratize the finance --- unique users/ transaction amt
Facebook- build community -- ads revenues/DAU/MAU
Cons:
Difficult to test or unable to track the difference
Long-term

Important KPI metrics
(1) Task completion: % of users who came here complete their primary task(order..)
(2) Share of search(paid search)
(3) Visitor loyal and frequencies(engagement and retention)
(4) Subscribers

(5) % of valuable exits(users clicking something value of you and leave)

2.2 Driver Metrics

More sensitive and actionable

E-commerce:
Checkout abandonment (to fix elements for improving revenues)
Users cart abandonment (check what users see before abandon, what campaigns, what produts)
Days to purchase(how long takes one people to complete the purchase)
Average Order value

Non-ecommerce:
Visitors
Frequencies of users
Recency of users
Length of visit
Depth of visit

B2B:
Number of accounts
Number of downloads of guidance
Number of free samples requested
Number of complete video watching
Number of visit to detailed pages

2.3 Counter metrics

What make you do not want to use this feature

Cart abandonment rate
Unfinished orders(check out abandonment)
Bounce Rate
Exit Rate

2.4 GUARDRAIL METRIC

To make sure the experiment run smoothly
Organizational guardrail metrics: page loading latency; errors per page; client crashes
Trustworthy : trustworthiness of experiments; violation of assumptions
Why?
Randomization are different
- numbers
- t-test or chi-squared test

2.5 Standards

- simple
- clear
- actionable
- low-variance metrics

Short-term revenues

(3) Hypothesis

When you confirm the metrics, what hypothesis you want to make
What is hypothesis testing:
Use sample of data to test an assumption regarding a population parameter
Null Hypothesis: u1=u2
Alternative hypothesis: u1!=u2

X1,....Xn ~ Bernoulli(p)

Bernoulli deals with the outcome of the single trial of the event, whereas Binomial deals with the outcome of the multiple trials of the single event.

Binomial = N Bernoulli adds up

P value: Given the null hypothesis is true, what is the probability that observing the extreme cases
Z value: critical values

T-test assumptions:

- Normality (if violated, just do it; transformation; other nonparametric methods)
- independence
- equal variance(two sample test)

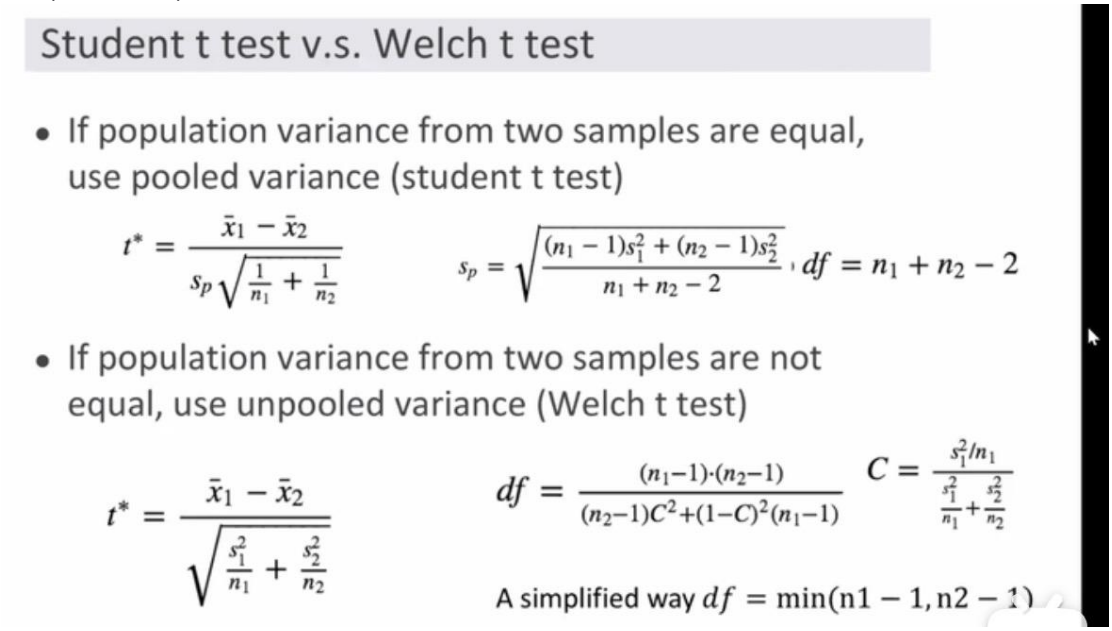
Degree of freedom

If population variance are equal:

n1+n2-2

If population variance are not equal:

Min(n1-1,n2-1)



3. Formulate the test plan

(1) Test

Because our population variance is not known

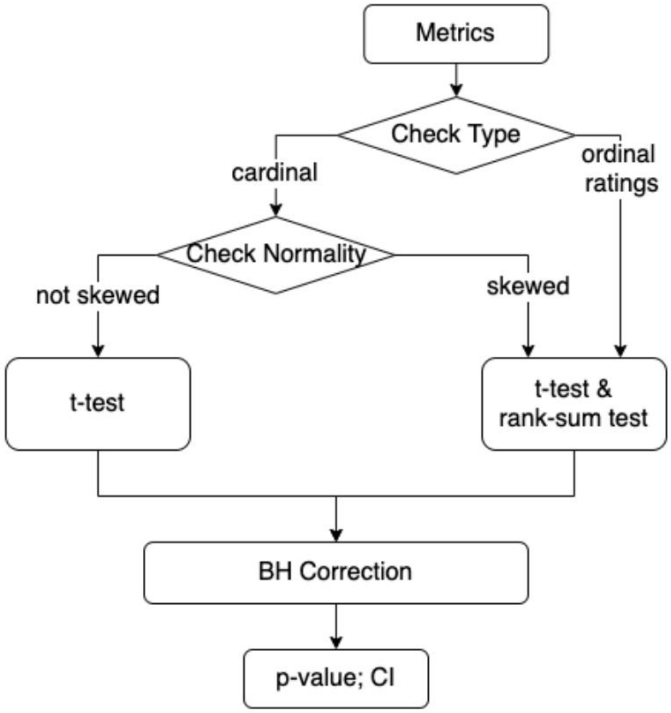
When we use the z-test for timespent A/B testing, we model the distribution as a normal variable, with mean $\mu = \frac{1}{N}\sum x_i$ and variance $\sigma^2 = s^2/N$, where $s^2 = \frac{1}{N}\sum(\mu - x_i)^2$. The problem is, we cheated a little: we used s^2 and not σ_p^2! We do this because we don't know σ_p^2, all we have is the estimate s^2.

The t-test models this uncertainty in the estimation of σ^2. When we perform a t-test, it feels very similar to the z-test, except in some places we write $N-1$ instead of N. And in the end, we don't look up a z value on a normal distribution, instead we look up a t value on a t-distribution:

8.1 Continuous metrics and ratios: T_TEST

8.2 Two bullet points that do not have a normal distribution -- Wilcoxon Rank-Sum test

8.3 Categorical variables: Chi-square



When variance is known (proportion)

- Z-test

When sample size is small use non-parametrics

When distribution is not normal:

Use Bootstrap

(2) Significance level, statistical power and practical difference

Sample size increase
-- SE reduces
-- beta decrease
-- statistical power increase

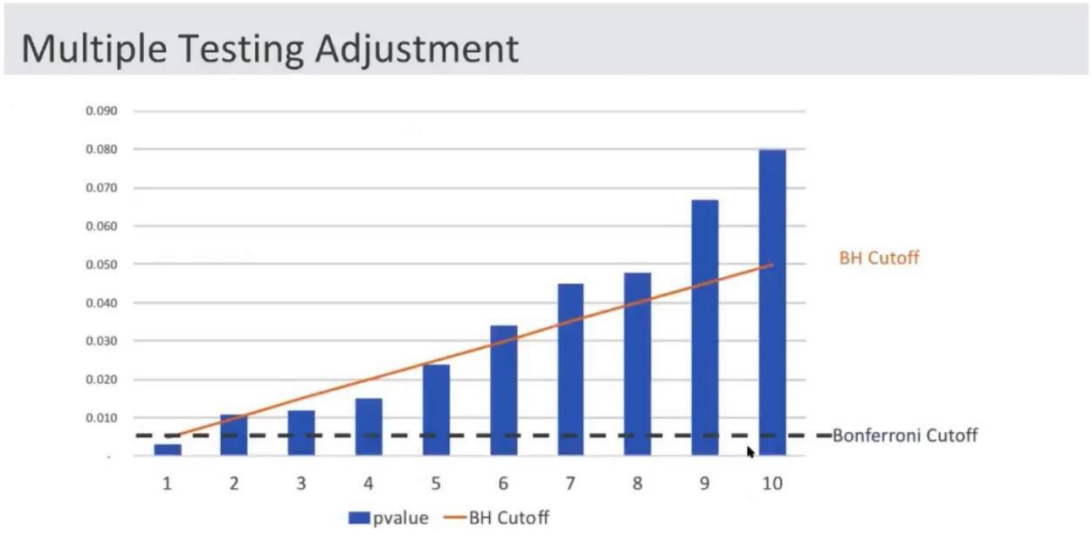
2.1 Multiple testing problem

Type I error (false positive). reject H0 when H0 is true
Type II error(false negative). accept h1 when h0 is false

Type1 error is more serious than type 2 error because lauching new wrong features might have more influence than insisting the old one

- simultaneous testing of more than one hypothesis
- Type I error may occur
1-(1-a)^k
With the k increase, the probabilities that type I error increase
More possibilities to get false-positive results
- A kind of data peeking

Way to deal with it
- Bonferroni correction a=a/k
The Bonferroni correction tends to be a bit too conservative and is based on the assumption that all tests are independent of each other.
- Redo the experiment with only one variant(more easily to explain)
- Fasle discovery rate



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

False discovery rate= #false positives/rejects
5 FP 95 TP FDR=5/100=0.05

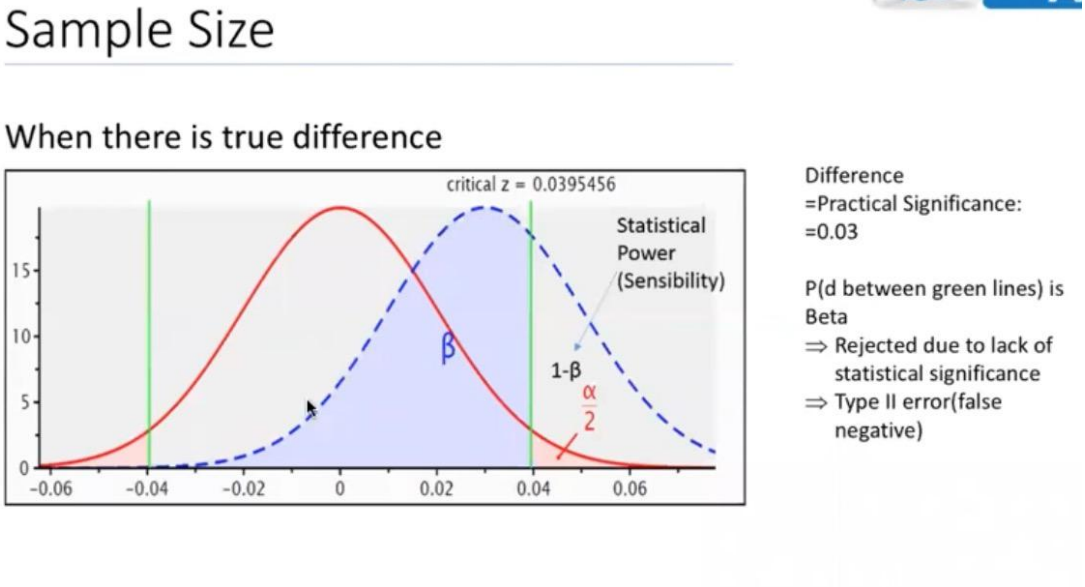
- two-step rule of thumb

- separate all metrics into three groups
    - those you expect to be impacted a=0.05
    - those potentially to be impacted a=0.01
    - those unlikely to be impacted a=0.001
- apply tiered significance level to each group

2.2 Lack of statistical power

- test may be under-powered to detect the effect size
- not enough randomization units in the test
    - less users than desired

What is statistical power?
statistical significance is how likely it is that the difference between your experiment's control version and test version isn't due to error or random chance.



4 factors might affect statistical power
4.1. Sample size
4.2. Minimum Effect of Interest (MEI, or Minimum Detectable Effect)
The Minimum Effect of Interest (MEI) is the magnitude (or size) of the difference in results you want to detect.
4.3. Significance level (α)
4.4 Desired power level (implied Type II error rate)
With 80% power, you have a 20% probability of not being able to detect an actual difference for a given magnitude of interest.

Conclusion
Statistical power helps you control errors, gives you greater confidence in your test results, and greatly improves your chance of detecting practically significant effects.

Take advantage of statistical power by following these suggestions:

1.Run your tests for two to four weeks.
Time= sample size/number of
2.Use a testing calculator (or G*Power) to ensure properly powered tests.
3. Meet minimum sample size requirements.
4.If necessary, test for bigger changes in effect.
5. Use statistical significance only after meeting minimum sample size requirements.
Plan adequate power for all variations and post-test segments.

Balance learning effect and user experience

Rolling-out plan

5%-10%-20%-50%...100%

(3) Sample size

3.1 Formulates

$$\frac{16 * \sigma^2}{\delta^2}$$

N=2σ^2(Zβ+Zα/2)^2/difference^2

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

(assume equal sized groups sample size in each group)

Zβ desired statistical power

Zα/2 statistical p

σ^2 desired outcome variance ---estimate(sample variance from AA test)

Difference(observational data; qualitative result; minimum effect worth the change)-- what if real difference is low, reestimate the sample size

The rule of thumb is that sample size n approximately equals 16 (based on a= 0.05 and b= 0.8) multiplied by sample variance divided by δ^2, whereas is the difference between treatment and control:

σ variance- can come from others' experiments

Actually, we don't know δ before we run an experiment, and this is where we use the last parameter: the minimum detectable effect. It is the smallest difference that would matter in practice

Power analysis:

P(x>1.96| u=3,σ=1)= P(z>1.96-3/1)=85%

3.2 Mistakes

1. Biased Sample
Business cycles; promotions; ad campaign;
We can use A/A experiment to see whether it is biased
2. too small sample
Calculate the minimum sample size
Type I error happens when we reject the null hypothesis when it should not be rejected. A common value for alpha is 0.05.
Statistical power is the probability that the test rejects the null hypothesis when it should be rejected.(0.8)

$$n = \frac{(Z_{\alpha/2}\sqrt{2p_1(1-p_1)} + Z_{\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)})^2}{|p_2 - p_1|^2}$$

p1 is the "Baseline conversion rate"

p2 is the conversion rate lifted by Absolute "Minimum Detectable Effect", which means p1+Absolute Minimum Detectable Effect

α is the "Significance level "

β is the in "Statistical power 1−β"

z a/2 means Z Score from the z table that corresponds to a/2

Zβ/2 means Z Score from the z table that corresponds to β/2

(4) How long to run an A/B test?

Minimize the exposure and duration of an A/B test
- optimize business performance
- potential negative user experience
- inconsistent user experience
- expensive to maintain multiple versions

sample size/# of users

- minimum sample size

- daily volume & exposure %

- seasonality(day of week)

we can obtain the number of days to run the experiment by dividing the sample size by the number of users in each group. If the number is less than a week, we should run the experiment for at least seven days to capture the weekly pattern. It is typically recommended to run it for two weeks.

- More than 1 treatment group?

not same significant levels.

significant.(1-0.05)^3

the False positive will increase(Type I error)

How to decide exposure

- size of eligible populations
- Potential impact
    - user experience
    - business impact
    - easy to test/debug

What if it takes to long to get desired sample(variance too much)

- increase exposure
- reduce variance to reduce required sample size
    - Blocking -run experiment within sub-groups
    - Propensity score matching

        1. run a model to predict CTR with appropriate covariates
        2. Check the propensity score is balanced across test and control groups
        3. Make 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

(5) Unit of diversion

Cookies or User-id Device_id,session_id, IP address
Would it be okay that users log-in different time than only one time to achieve their targets?
- what are the eligible subjects we try to influence
    - Not registered yet (cookie) session_id will influence users' experience
    - A fixed group of users
- what is the objective
- independence & user experience
    - Example: change homepage design in an app(session_id)
        Add new video chat filters(cluster_id...network effects)

Spilt %-% of users in test/control
50%-50% might lose best opportunity time(time sensitive & holiday marketing campaign)
If control group is costly,marketing campaign

80%-20%: imbalanced data; unequal variations(not applicable for T-test)

Problem:
Independence:
    -Non-login users: assign by cookie_id;device_id
    -Multi device users
    -Multiple users share one device: predict users with model
Randomness:
    Deterministic assignment: AA test
Reproducible:
    Set salt: user_id fixed in one group after first
50/50 split:
    Imbalance assignment: Understand why? Change assignment methods?
Test/Control are comparable

(6) Randomization and Variant

-invariant checking(sanity check): metrics shouldn't change between your test and control
- same number of users
    How to confirm:
    A/A test
        Make sure the framework been used is correct
        Data exploration & parameter estimation(sample variance)

Determine users are randomly splits
The results are non-significant all the time

- is the distribution same? (normal distributions)

If not concerned with daily effects,use aggregated metrics; daily metrics would increase the sample requirements(because it is more variant)
Usually two-groups
If multivariate: do not affect each others; user Bonferroni corrections
Longer to detect

4. Run the experiment

4.1 Can we do a peeking?

- We could not simply stop it because the p value <0.05 because the n is preset and stop earlier would increase type 2 errors
- When can we do that
The new test harming users
The change the statistical power

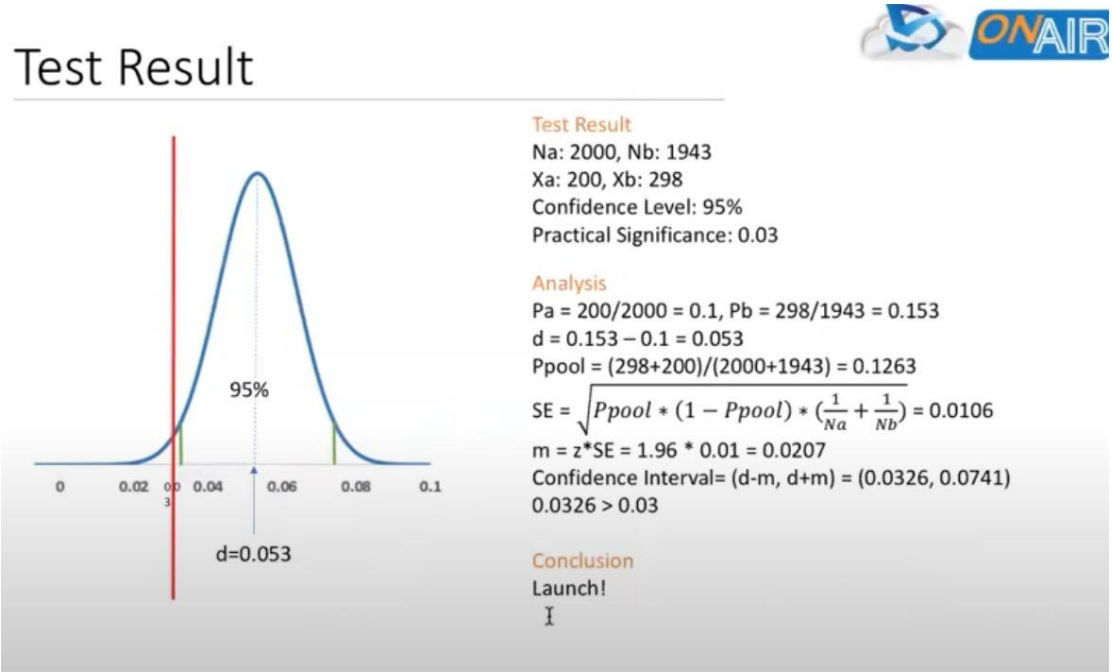
5. Analyze the result

5.1 Data Exploration

- Imbalance assignments
    A/A test(randomization assignment)

- Mixed assignment
    If only little of users are mixed, delete?
    However, users might be highly-active(multiple device) that they will be mix-assigned
    Might only consider it as test(dilute your test effect)
- sanity check
    Other factors?

(1) P<0.05



1.1 What affects?

Meet expectations?

- statistical power
- significance level
- day of week effect
- seasonality
- novelty effect and change adversion

If result is too good?
May be the result of outliers

1.2 Launch or not?

--- depends on the trade-off computation cost,labor cost, users adaptation...
--- filtering all conditions
--- maybe not repeatable (seasonal effects/ change aversion/ holidays)

1.3 Next steps

- Users experience research
- Focus group
- Cohort analysis
    How to measure impact over time?
    Select a cohort of users and monitor their metrics change over time
- Survey

1.4 If two metrics conflicts?

“Choosing between A or B depends on which has a higher impact on their common metric for each different segment.”
Define a common metrics
Maximum the revenues
Goals of the team
Are they expected? Are these metrics important?
Deep dive to find causes

(2) P>0.05

2.1 Any external effects?

- Competitors
- Industry change
- Other experiment effects()

2.2 Internal effects

- technical issues
- other promotions
- maybe divide into different segments of users)
    Statistical reasons for counter-intuitive results : Simpson’s paradox
    Within the groups significant
    Aggregate not significant
- A/A test
    See any pre-bias

6. Other Questions

(1) How to change the predicted practical change in metrics based on KPI metrics

	Impact to KPI Metric in Test	**% of users impacted by test (e.g. wallet users/ all app open users)**	**Impact to KPI Metric Topline**
	+x%	y%	+x%*y%
Example: Making improvements to the Android onboarding flow	+2% lift in First Trip Conversion observed in the test	20% of user-base is on Android	+0.4% aggregate topline impact to First Trip Conversion

(2) Simpson dilemma

By different segements, group A successful rate always higher than GroupB
Combining together, group A successful rate is lower than groupB
Reason:
Different segment, the successful rate had a significant difference
Group B apply more resources into high successful rate than GroupA

(3) PRE-Bias

Regression adjustment

Y_post=beta_pre* Y_pre + beta_t* Treatment_group

Diff- in - diff comparsion

(Y_post/t-Y_pre/t)/(Y_post/c-Y_pre/c)