BUSINESS ANALYTICS

商业分析

Data Application Lab

AB testing

“Most people are wrong most of the time”

—Anonymous

Outline

1. A/B Testing Overview
2. A/B Testing Procedure
3. Case Study – da-shu-ju.com
4. Interview Questions

A/B Testing Overview

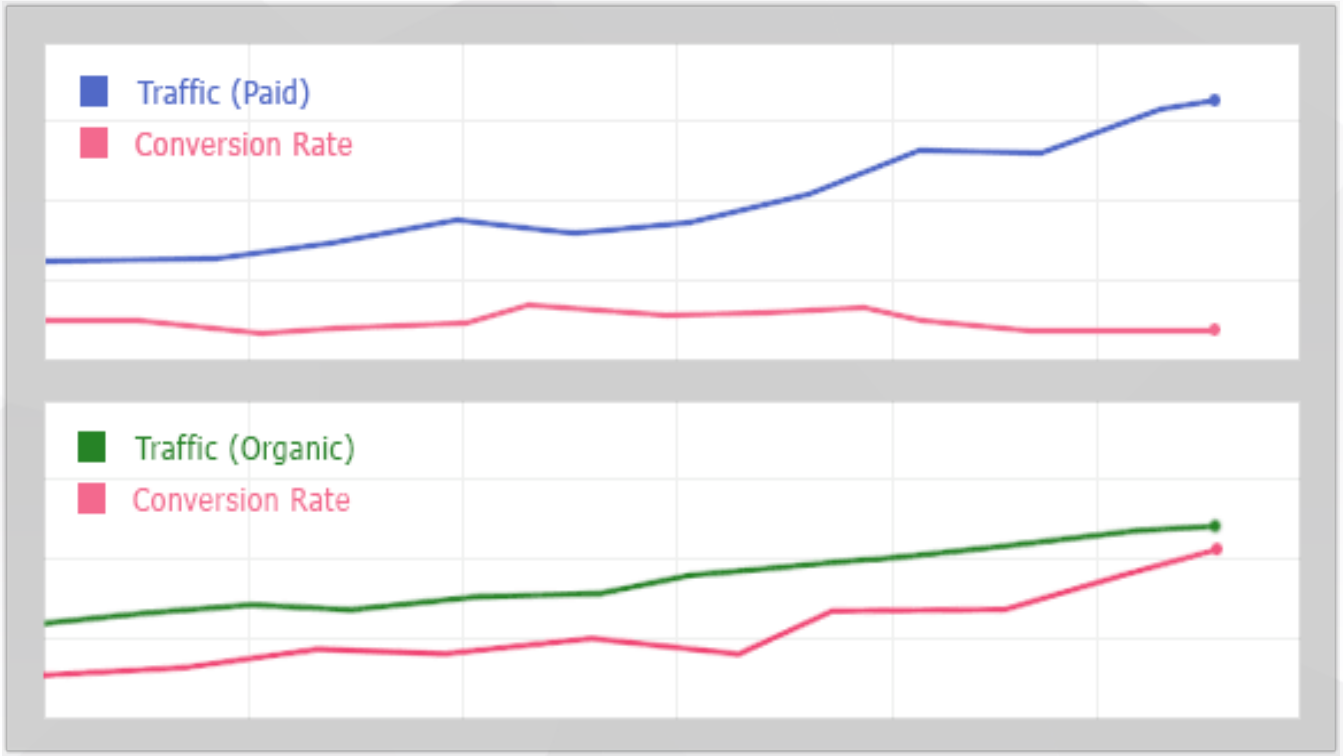
What is A/B Testing

Experimentation is the engine that drives innovation. Experiments show us, with scientific rigor, what works and what doesn't work

-- PROF. STEFAN THOMKE, HARVARD BUSINESS SCHOOL

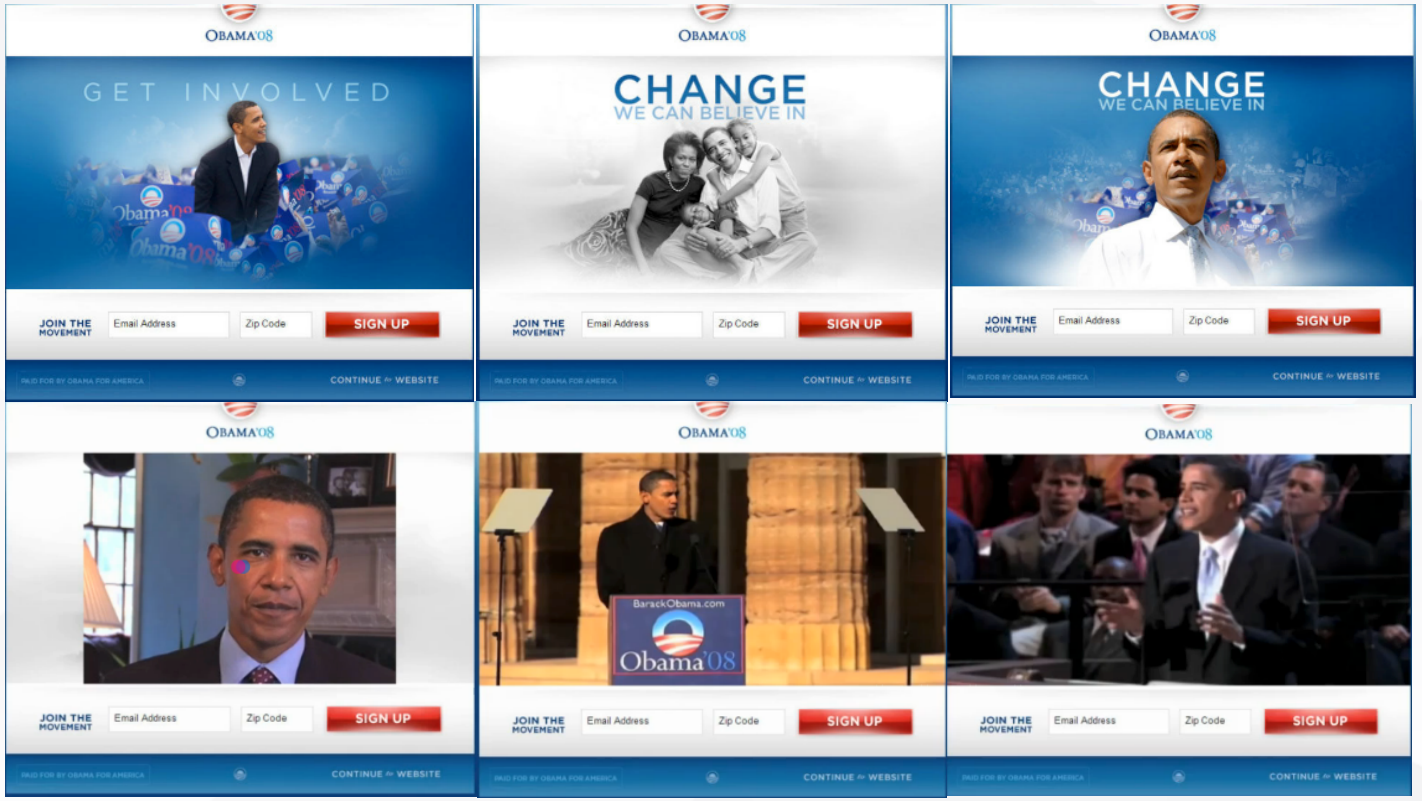
Why A/B Testing

- Effective way to find out the truth even before the Internet exists
- Most people are wrong most of the time
- **Cost effective to grow conversions organically!**
- Reduce Risks
- Reduce bounce rates & increase engagement



How Obama Raised $60M More Thanks to A/B Testing

4 X 6 Test Versions Created



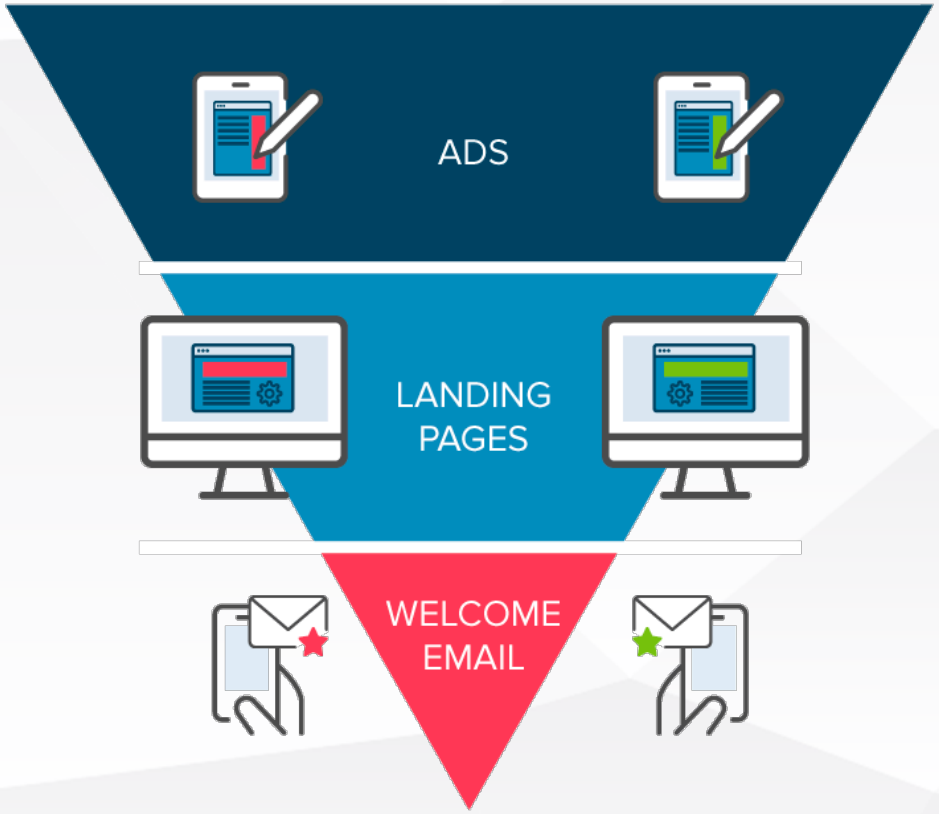
Test Results Compared

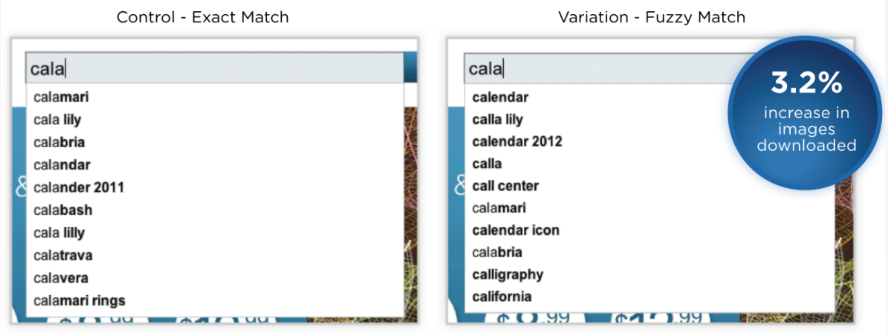
Combinations (24)

CSV | TSV | Print

Relevance Rating	Variation	rved t	Conv./Visitors
Button	Original	—	5851 / 77858
5 / 5	Learn More	.6%	6927 / 77729
	Join Us Now	37%	5915 / 77644
	Sign Up Now	38%	5660 / 77151
Media	Original	—	4425 / 51794
5 / 5	Family Image	.1%	4996 / 51696
	Change Image	85%	4595 / 51790
	Barack's Video	14%	3992 / 51427
	Sam's Video	.4%	3261 / 51864
	Springfield Vid	.3%	3084 / 51811

Where to A/B Test?





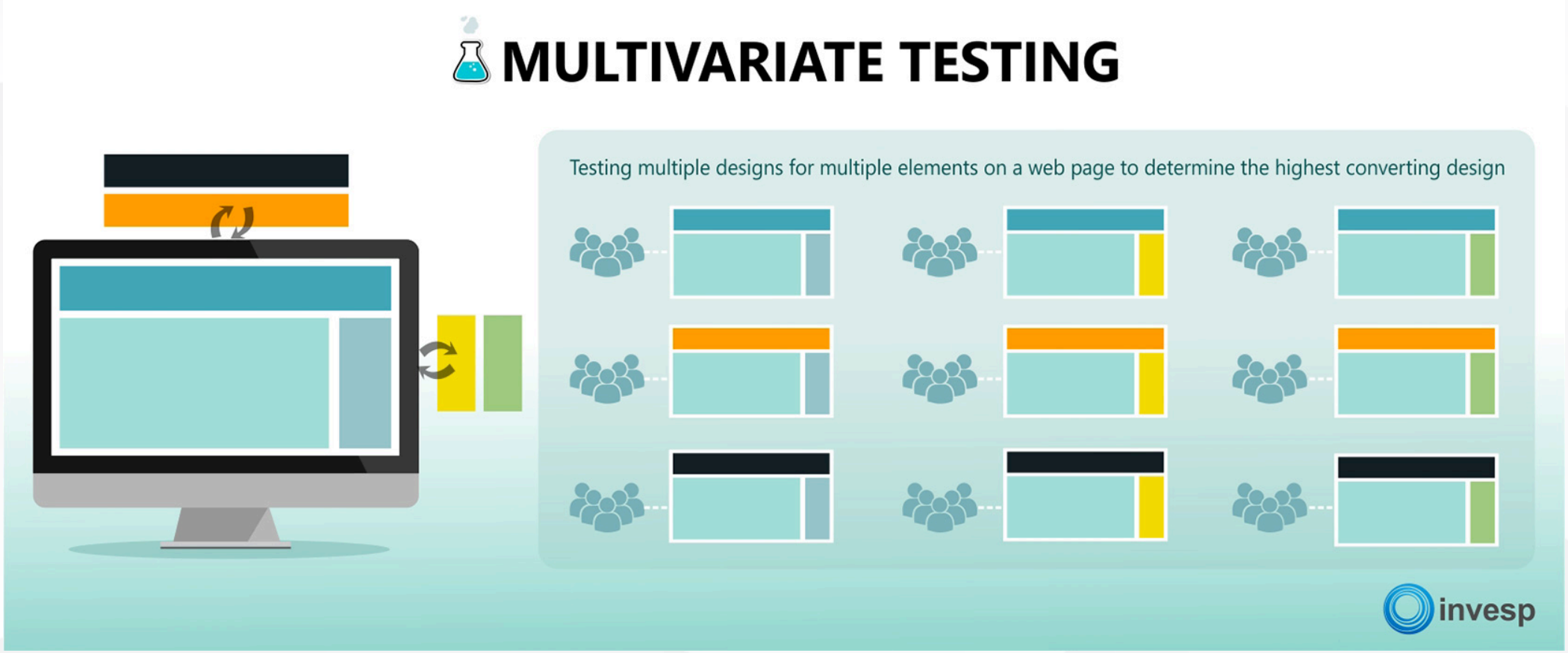
Where not to A/B Test

- No clear comparison between control and experiment group
- Emotional changes needs time: Logo/Brand Name
- Response data hard to gather
- Too long to response

Quiz

- Which of the followings can be evaluated by A/B testing?
  A. An online training website to launch premium service ✗
  B. A retail brand to apply new customer segmentation for email marketing ✓
  C. A social media site wants to know if changing the color of “Sign up” button helps with conversion ✓
  D. A e-commerce site wants to know if they are missing some good selling-products ✗
  E. A career service site wants to know if the new job recommendation algorithm can improve hiring rate ✗
  F. Video sharing sites wants to test new video recommendation algorithm ✓
  G. Company wants to change brand name ✗

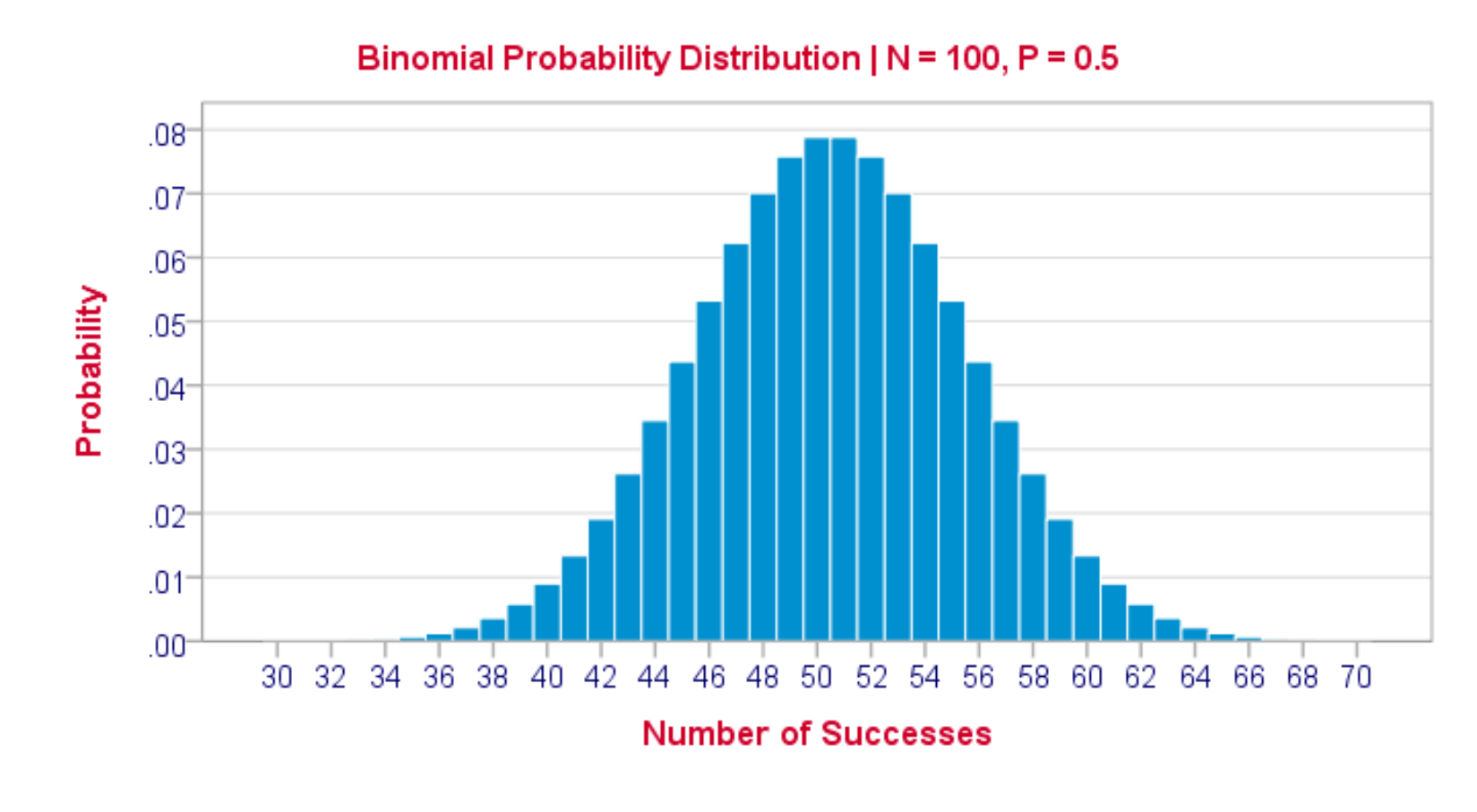
A/B/C Testing and Multivariate Testing



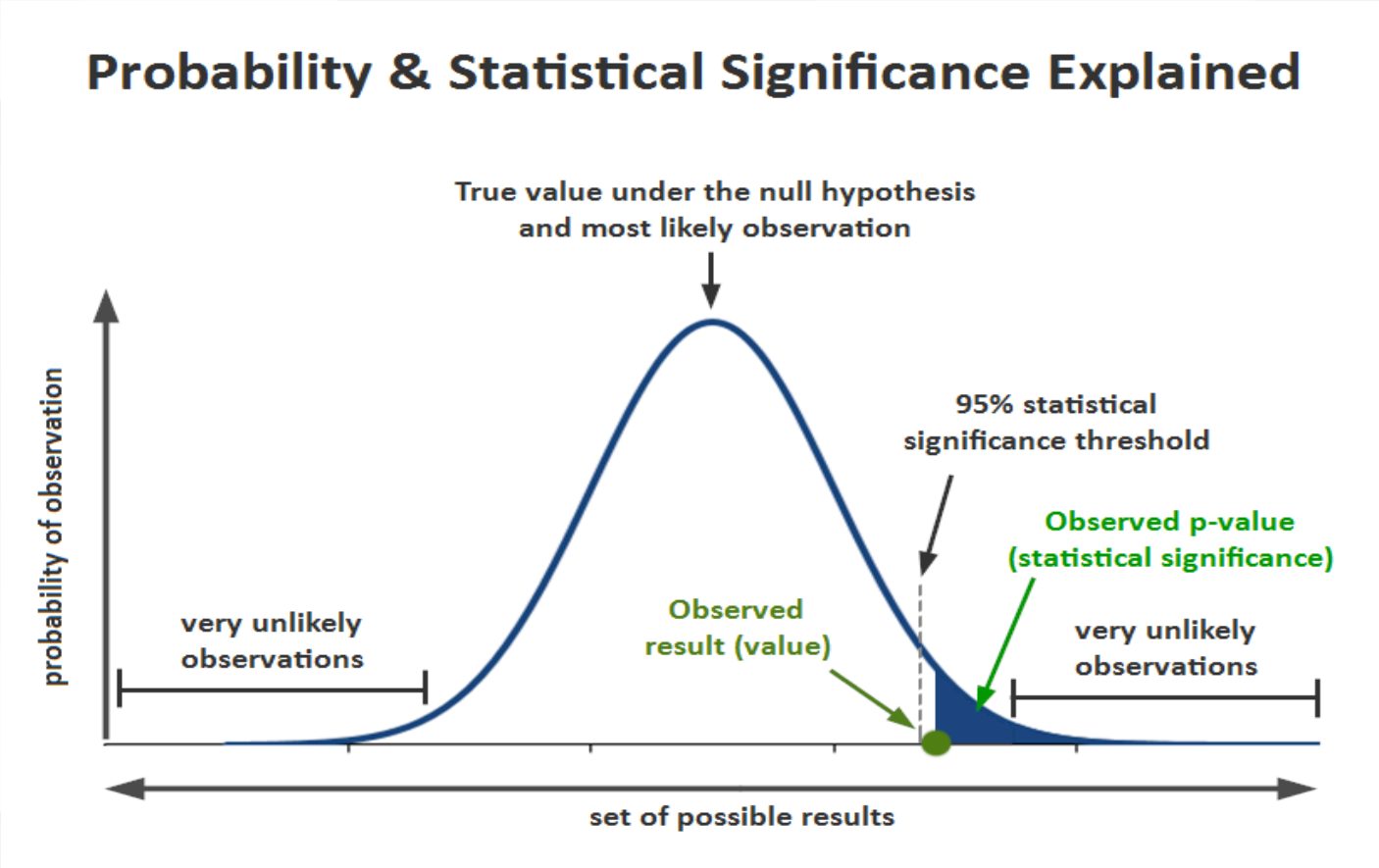
A/A Testing

- Test two groups of identical pages
- Expect the test results to be very similar or no statistical significance
- Ensure the experiment setup is correct
  - Misreporting of key performance indicators
  - Landing page display problems
  - Stopping a test too early
- Often run before A/B Testing

Binominal Distribution



Null Hypothesis & Statistical Significance



Measure Result

Group	Control-A	Variation-B
Unique Visitor	500	500
Click	50	60
Conversion Rate	10%	12%

Is B the winner?

Is the difference statistical significant?

- H0 : d0 = 0
- d' ~ N(0, SE)
- Ppool = (Xa + Xb)/(Na+Nb)

  =(50+60)/(500+500)

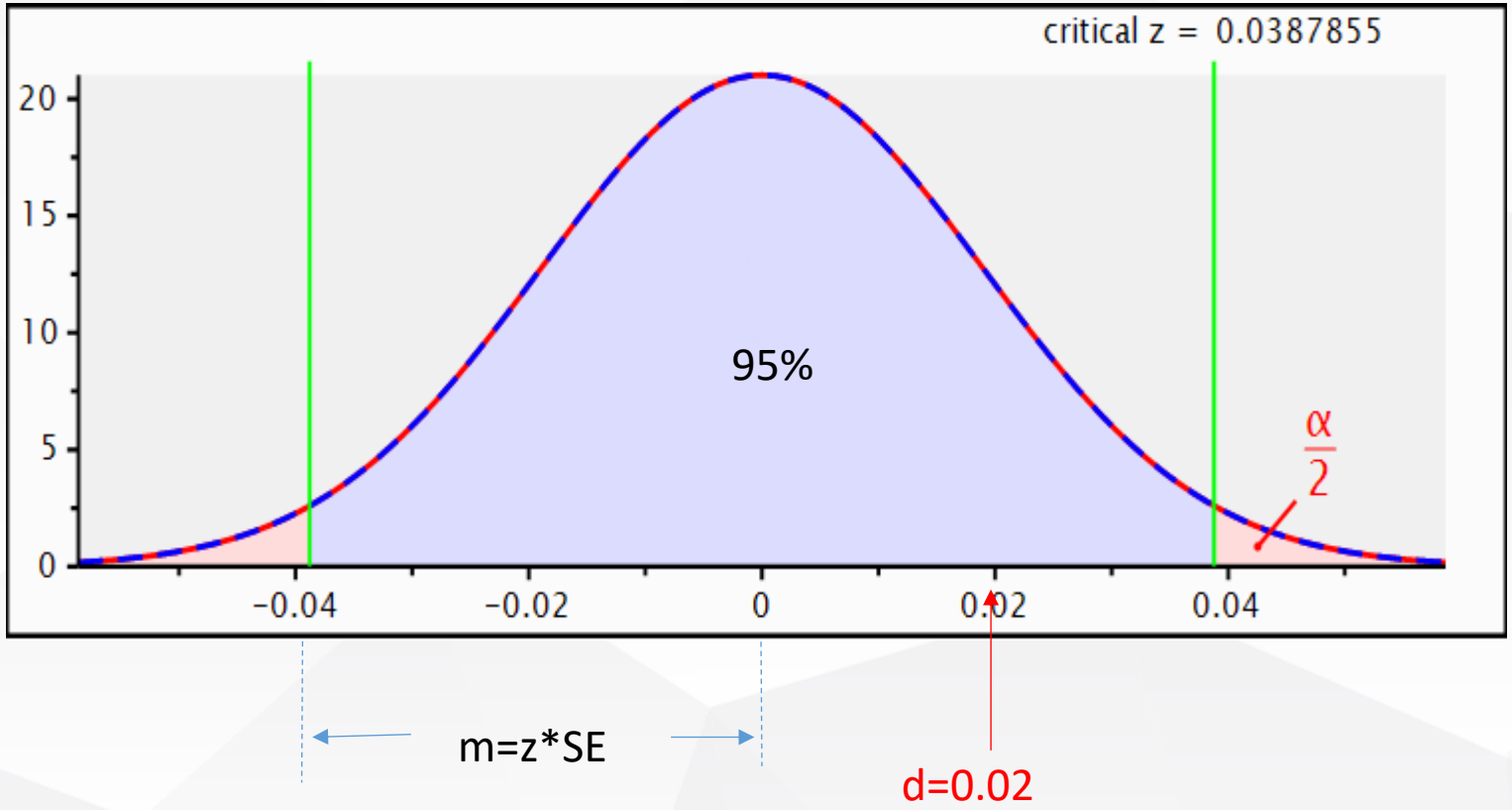
  = 0.11
- SE = $\sqrt{\frac{p(1-p)}{N}}$

  $= \sqrt{Ppool * (1 - Ppool) * (\frac{1}{Na} + \frac{1}{Nb})}$

  $= \sqrt{0.11 * (1 - 0.11) * (\frac{1}{500} + \frac{1}{500})}$

  = 0.019789
- 95% confidence level (Two tailed)=>z=1.96
- m = z*SE = 1.96* 0.019789 = 0.0388
- Confidence Interval: (-0.0388, 0.0388)
- d = 0.02 between 95% CI

=> Not statistical significant!



What if sample size is 4X bigger

- H0 : d0 = 0
- d' ~ N(0, SE)
- Ppool = (Xa + Xb)/(Na+Nb)
  =(200+240)/(2000+2000)
  = 0.11
- SE = $\sqrt{\frac{p(1-p)}{N}}$

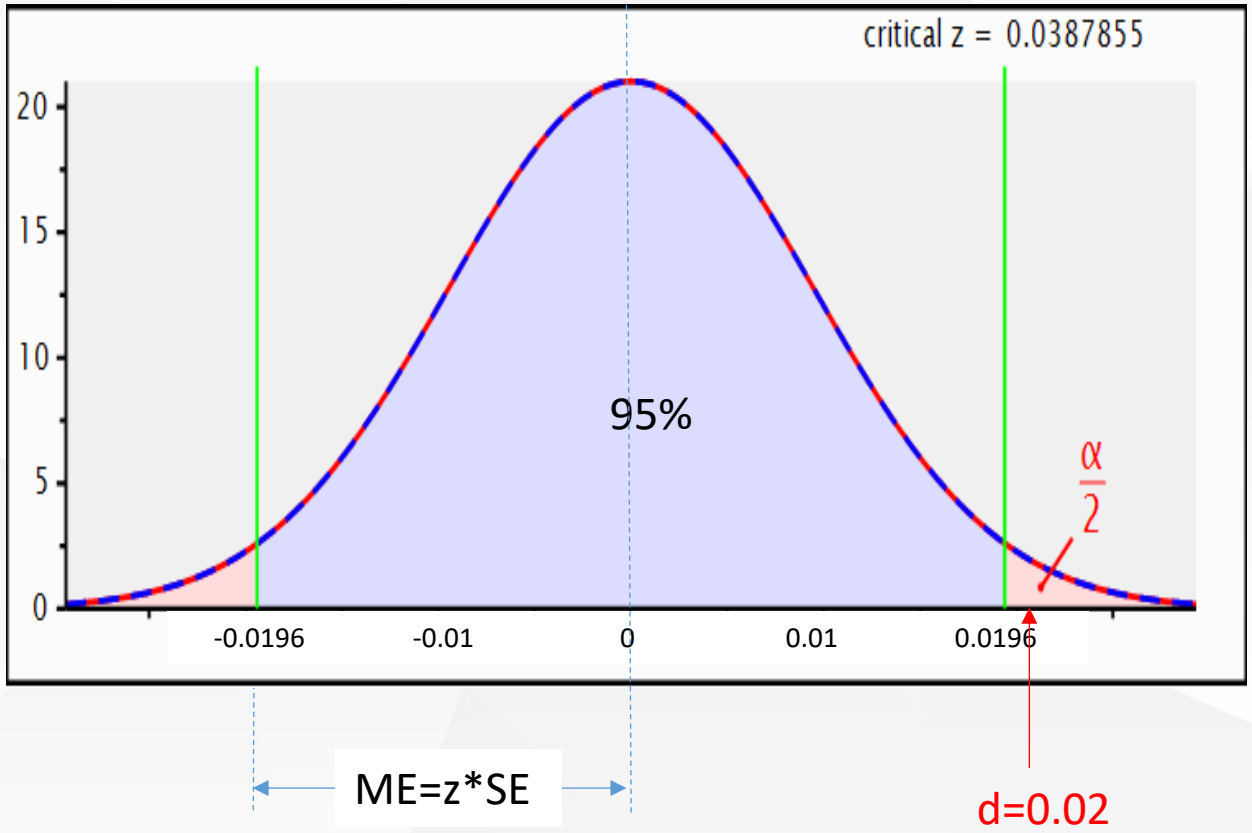
$$= \sqrt{Ppool * (1 - Ppool) * (\frac{1}{Na} + \frac{1}{Nb})}$$

$$= \sqrt{0.11 * (1 - 0.11) * (\frac{1}{2000} + \frac{1}{2000})}$$

= 0.01

- 95% confidence level (Two tailed)=>z=1.96
- ME = z*SE = 1.96* 0.01 = 0.0196
- Confidence Interval: (- 0.0196, 0.0196)
- d = 0.02 > Upper Bound CI

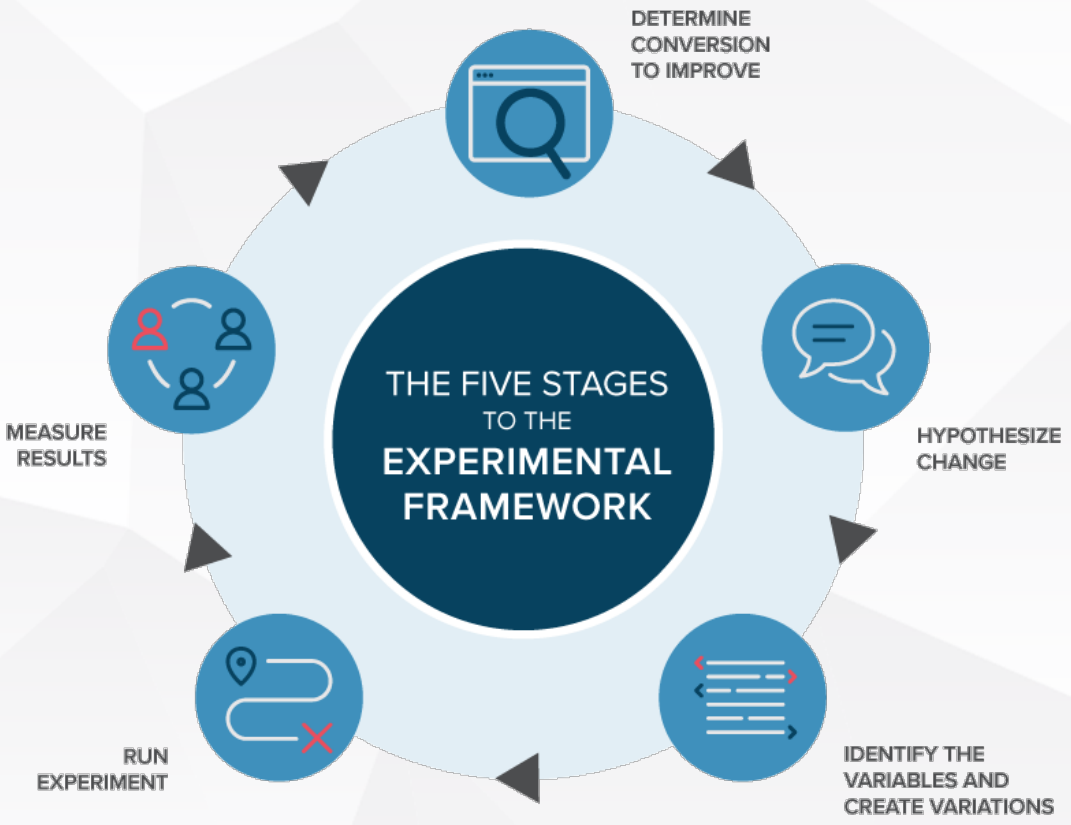
=> statistical significant!



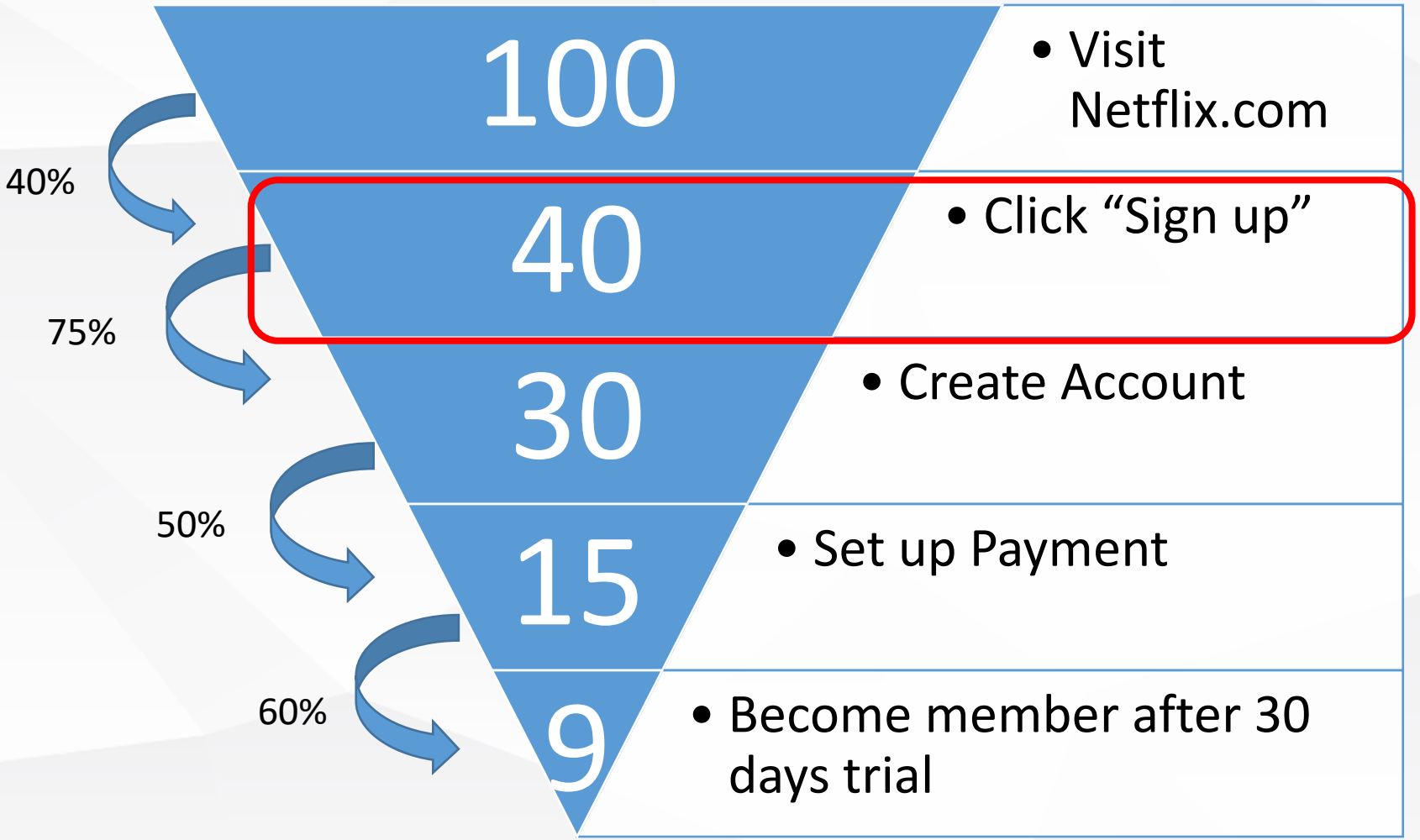
A/B Testing Procedure

A/B Testing Procedure

1. Review Business Funnel
2. Determine Process to Improve
3. Define Metric
4. Form Hypothesis
5. Design Experiment
6. Create Variation
7. Run Experiment
8. Analyze Test Result
9. Make Recommendation



Review Business Funnel



Common Web Analytics Metrics

- Count
  - Page View
  - Visit / Session
  - Click
  - Visitor / Unique Visitor
  - Active Users (DAU, MAU)
- Conversion
  - Click Thru Rate (CTR)
  - Click Thru Probability
  - User Click Probability
  - Bounce Rate
- Time
  - Active Time
  - Page View Duration
- Business
  - Revenue
  - Member
  - Order
- Ads Analytics
  - Impression
  - Click
- Composite Metric

Which metrics should you choose?

Invariant Metric

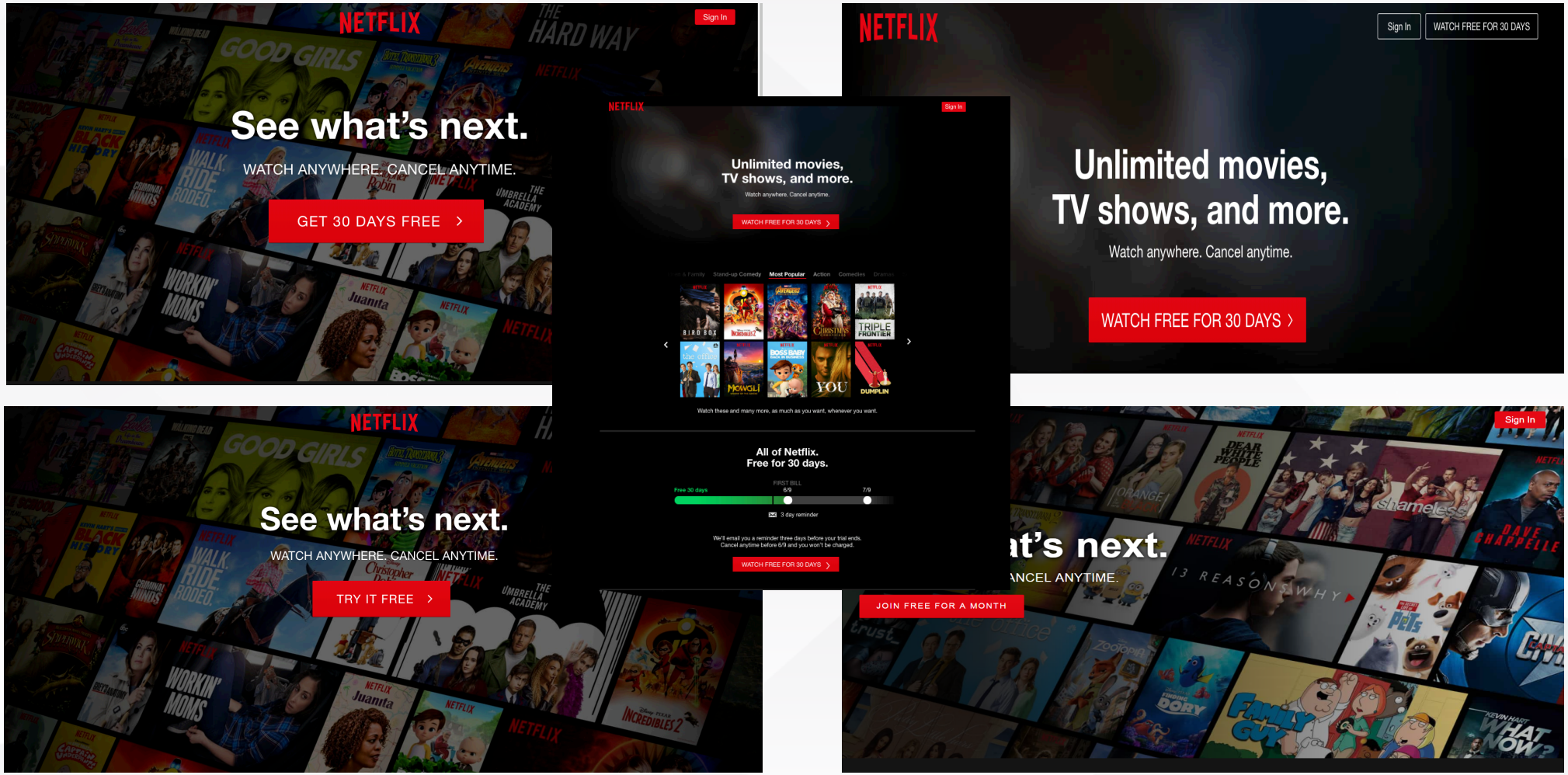
- Home Page PV
- Home Page Unique Cookies
- Home Page Visits
- Home Page Loading Time

Evaluation Metric

- Clicks
- Click Thru Rate
- Members/Unique Cookies
- Members/Clicks
- DAU

What can you change?

图片来源于网络



Hypothesis Change

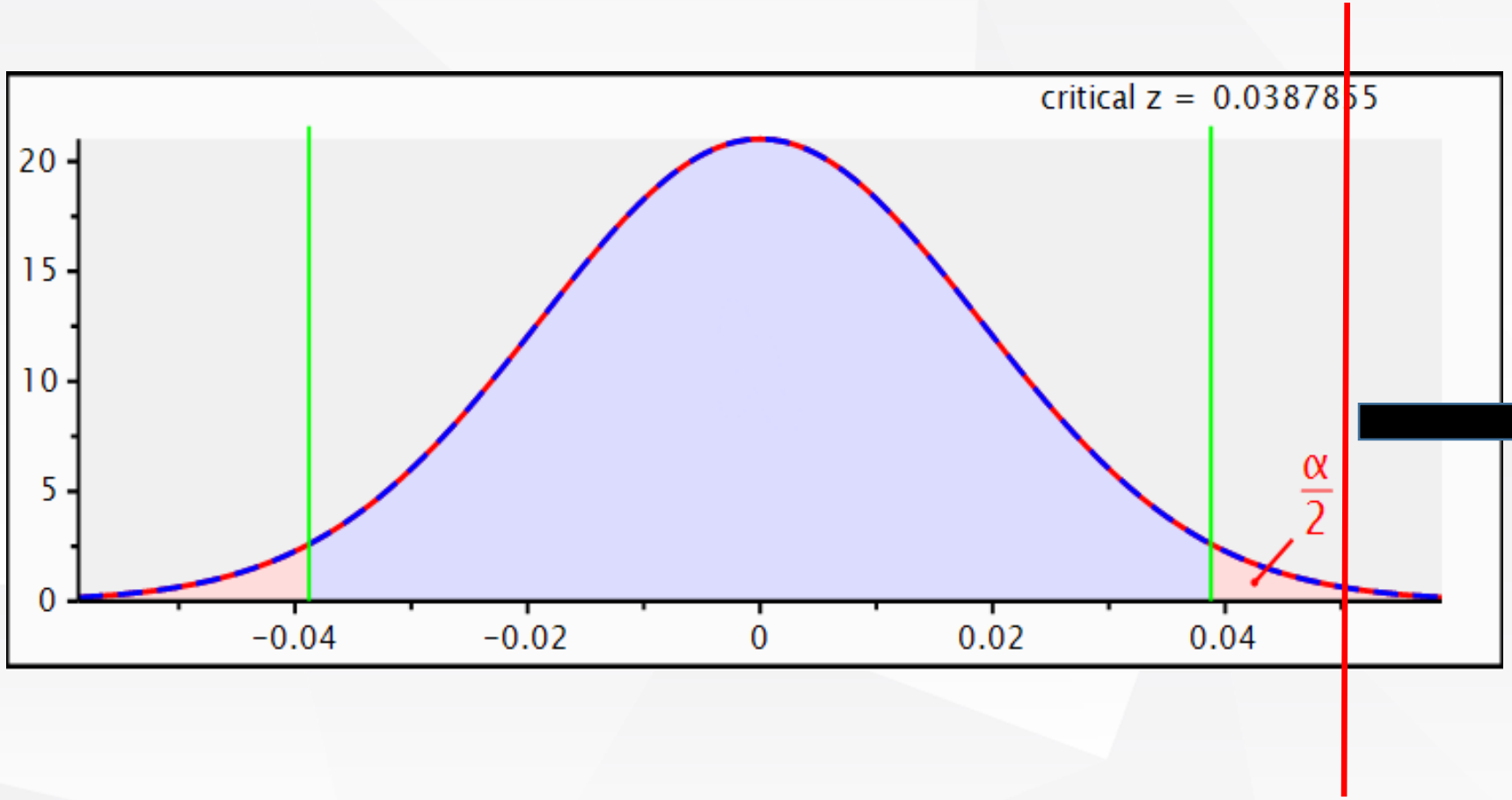
- Headlines
- Sub headlines
- Font Size
- Font Color
- Font Type
- Text Alignment
- Background Image
- Background Color
- Page Layout
- CTA Button Position
- CTA Button Color
- CTA Button Size
- CTA Text
- … …

Experiment Design

- What’s the business minimum detectable effect required to launch the change?
- How big should the sample size be?
- How long to run the experiment?
- How to divide the population randomly and evenly into A/B?
- What’s the target population for the experiment?

What's the minimum detectable effect?

- Also called Practical Significance
- The minimum effect should be obtained in order to launch the variance in production
- Change comes with cost. The lift from the change should at least cover the cost



Practical Significance: 0.05

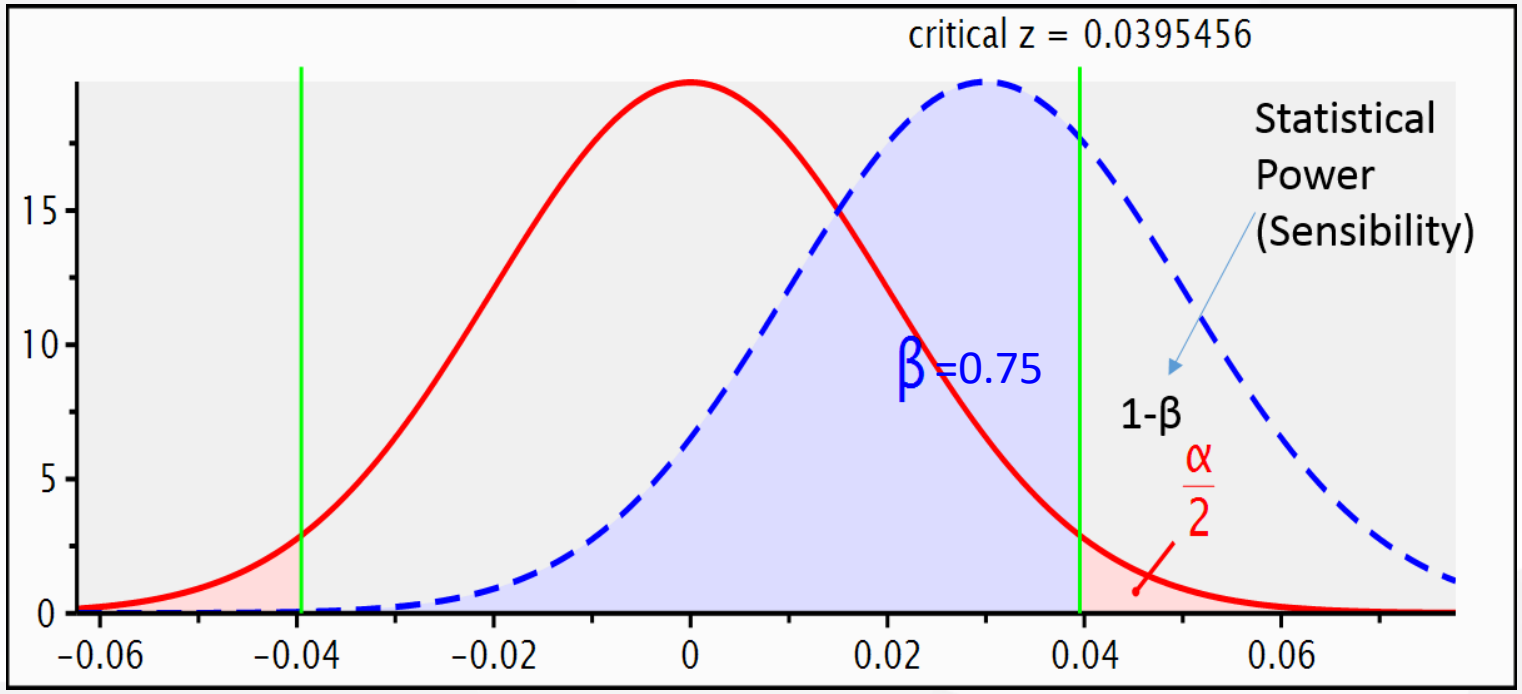
Quiz: Can you calculate the practical significance?

Company A issues credit cards to consumers. Direct mail is an important marking vehicle where company A sends letters to prospect customers with application form enclosed. Bank A wants to know if changing mail type from regular mail to priority mail will increase the response rate. The cost of a regular mail is **3** cents while the cost of a priority mail is **10** cents. Company A sends out **50,000** mails per year. The response rate by mail is approximately **0.1%** and approval rate is **50%**. An new credit card holder generates **$1,000** profit in average each year. How much should the response rate at least bump for Company A to consider the change?

Regular Mail	$0.03		Priority Mail	$0.10
# of Letters	50,000			
Original Respo	0.1%		New Response Rate	0.114000%
ApprovalRate	50%			
Profit/Card	$1,000			
Total Profit	$23,500.00		New Profit	$23,500.00

How should you choose sample size?

Practical Significance = 0.03. If there is a true difference of 0.03. N = 500/group.



If d(observed difference) fall into purple area

⇒ Null Hypothesis is not rejected

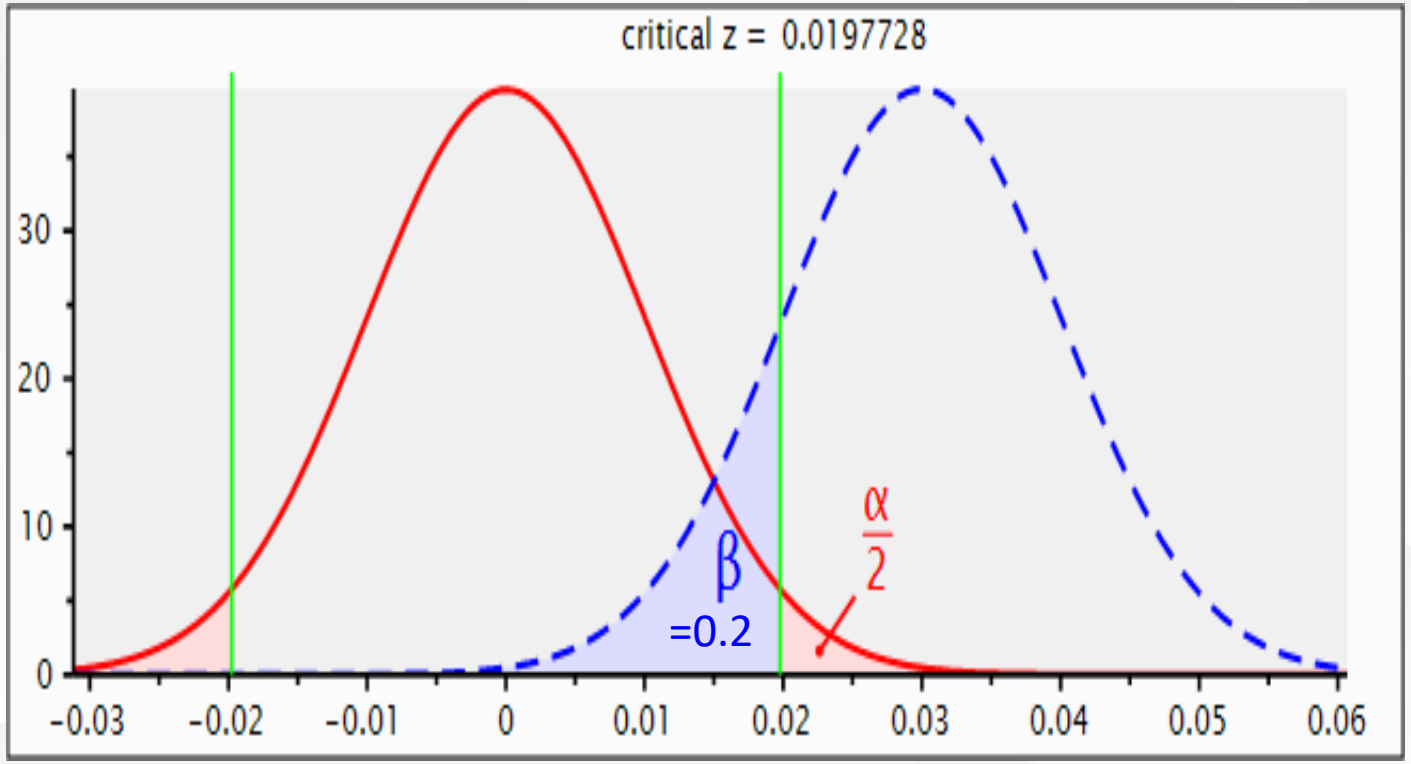
⇒ Lack of statistical significance

⇒ Type II error(false negative)

Sample Size and Statistical Power

N = 2000/group

https://i.gifer.com/DDjw.mp4



sample size Increases
⇒ SE Reduces
⇒ Beta decrease
⇒ Statistical Power(1-Beta) increase

General Practice:
Beta <= 0.2
Or Statistical Power > 0.8

Calculate Sample Size

1. Use Goal Seeker

N	1772.924769		Alpha	0.05
Pa	0.1			
d	0.03			
Pb	0.13			
P Pool	0.115			
SE	0.010714956			
95% Margin of Error	0.021000928			
Beta	0.200493217			

- [3. Use Sample Size Calculator]

2. Use Formula

$$n = \frac{2(\bar{p})(1-\bar{p})(Z_\beta + Z_{\alpha/2})^2}{(p_B - p_A)^2}$$

Z_β : z-score that corresponds to the level of statistical power

$Z_{\alpha/2}$: z-score that corresponds to the level of significance or confidence level

$\bar{p}$: pooled probability or average of p_A and p_B

p_A : success rate of control group

p_B : success rate of test group

	Value	Z-score
Beta	0.8	0.84162123
Alpha	0.05	1.95996398
Pa	0.1	
Pb	0.13	
Ppool	0.115	
N	1775.15497	

Quiz

- Which requires bigger sample size when following is required
  A. Higher practical significance
  B. Higher confidence level
  C. Higher statistical power
  D. Higher base click through rate (Assume CTR << 0.5)

How long to run your A/B Test?

- Too short: Bias
- Too Long: Bad user experience
- Common choice: 2-6 Weeks
- The higher the risk is, the lower to limit the traffic
  - Risk of losing customer/revenue
  - Conversion is time sensitive
  - Infrastructure change
- To decide duration: Sample Size, Daily Traffic, Traffic% for Testing

Example 1	
Sample Size/Group	1629
Unique Visitors/Day	500.00
Percent of traffic for testing	30.00%
Days to run the experiment	22

Example 2	
Sample Size/Group	15,000
Unique Visitors/Day	1,000,000
Percent of traffic for testing	0.1%
Days to run experiment	30

Test Duration Calculator

What’s the experiment target population

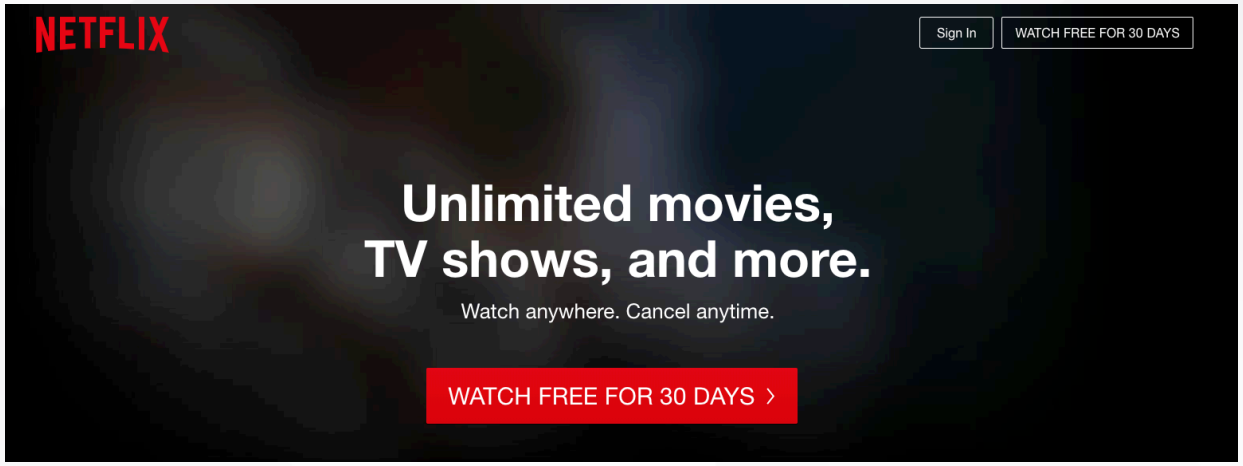
- Platform:
  - Web/Tablet/Mobile
  - Windows/Mac
- Browser:
  - Chrome/IE/Safari
- Language:
  - English/Chinese
- Geography:
  - China/US
  - State/Province
- Population:
  - Male/Female
  - Age
- User Segment:
  - VIP/Other,
  - Experienced/New
  - Level

How to divide the population into A&B?

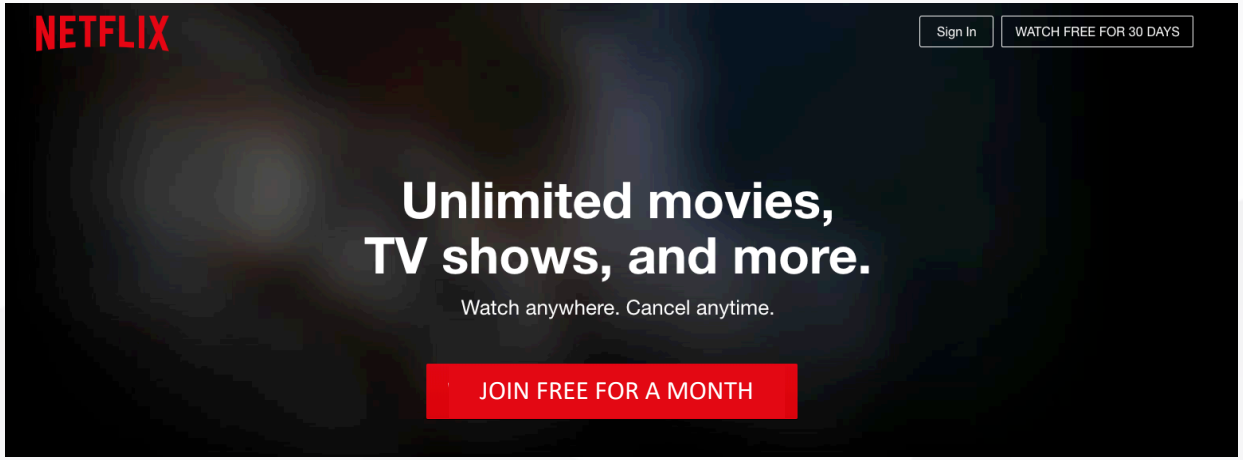
- User id
  - Unchangeable
  - Only available after a user signs in
- Cookie
  - Changes when changing browser or device
  - Can be cleaned
- Event
  - Randomly assigned when event triggered
  - No consistent experience
  - Use only for non-user-visible changes
- Device ID
  - Only available for mobile device
  - Changes when changing device
- IP Address
  - Changes when changing network

Identify Variable and Create Change

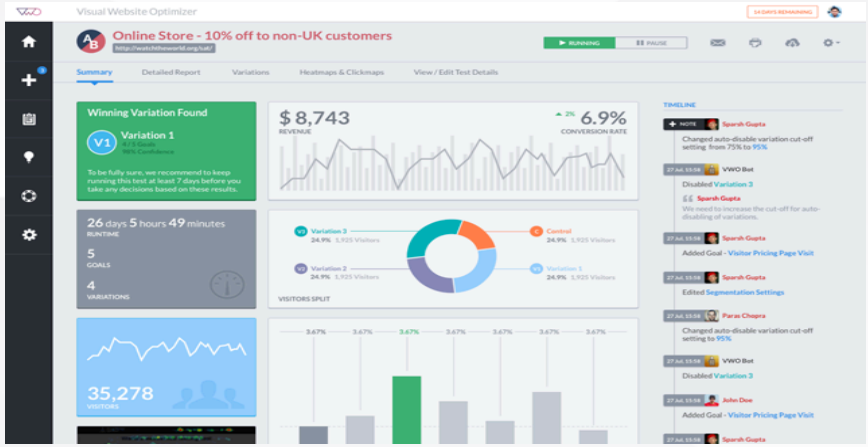
CONTROL



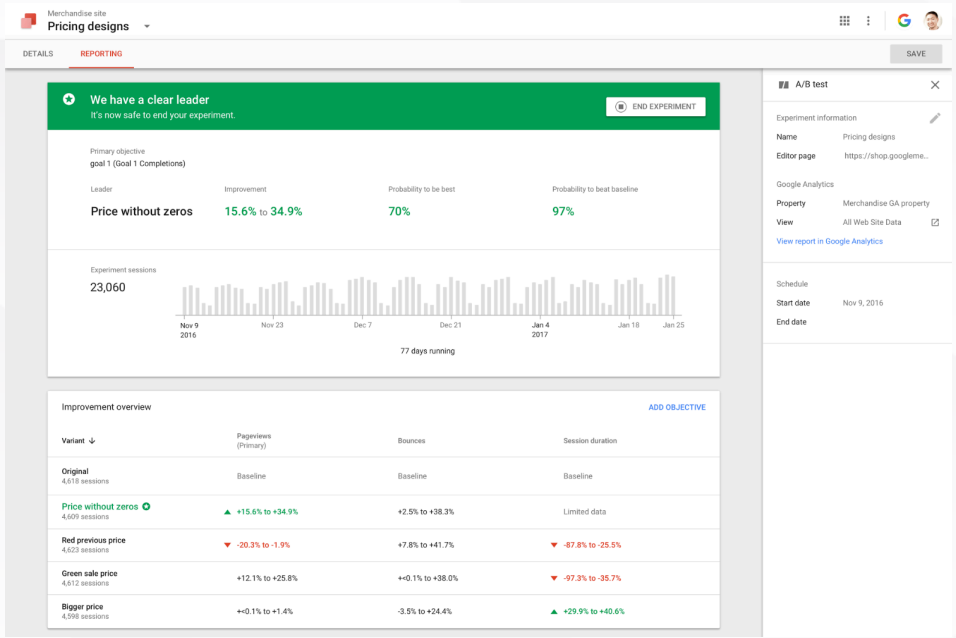
VARIATION



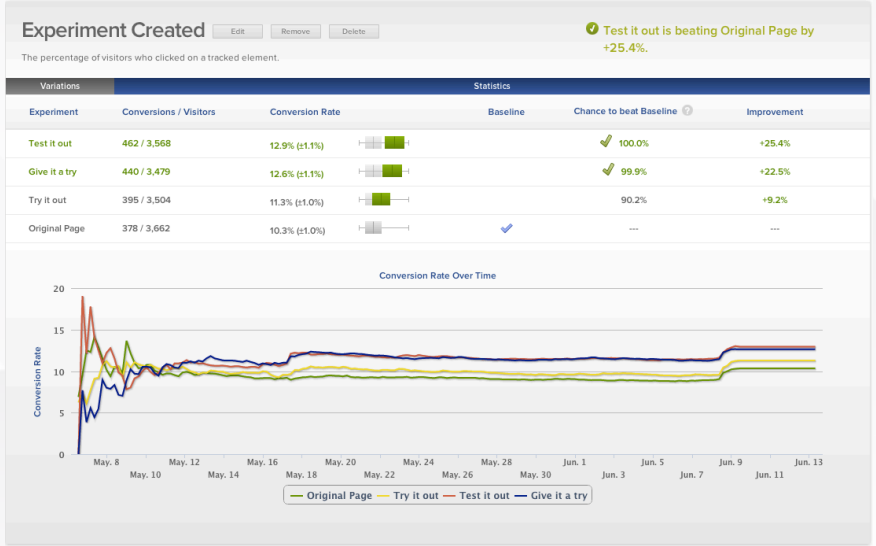
Run Experiment



Visual Website Optimizer (VWO)



Google Optimize



Optimizely

A/B Testing in Practice

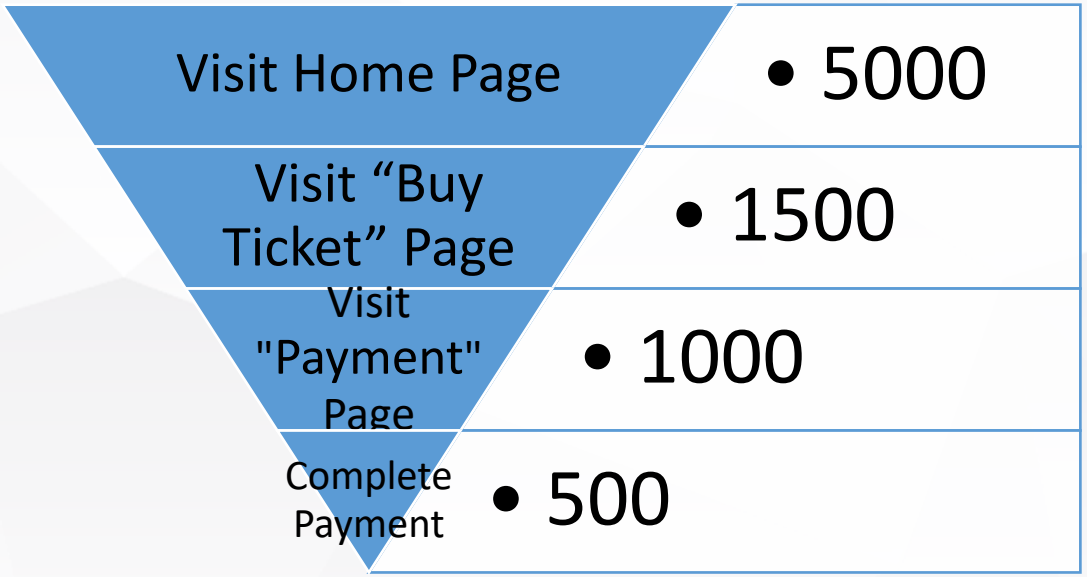
Case study: DA-SHU-JU.com

Case Overview

- Web Designer Conference 2019 sells tickets on da-shu-ju.com
- As a business consultant, you are given a task to boost the ticket sales as much as possible
- Main marketing channel include email and social media
- 3-level Price
  - Basic: $190
  - Standard: $289
  - Premium: $490

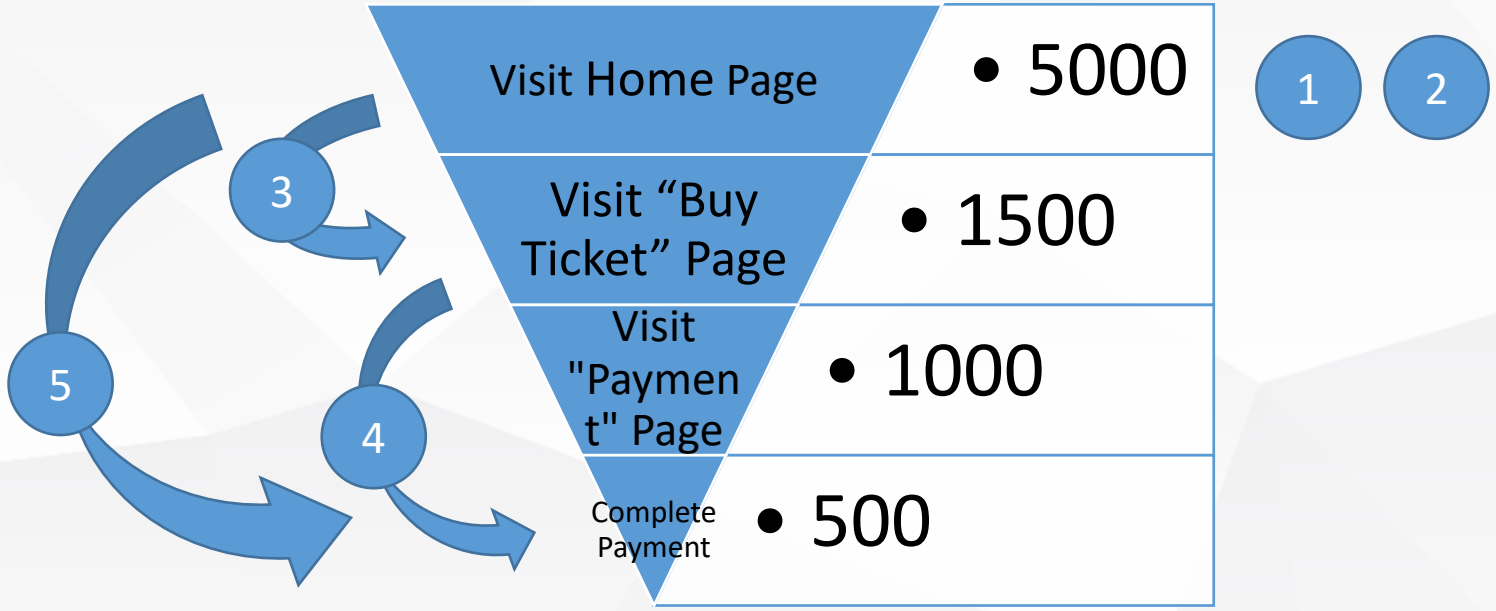
Review Business Funnel

Funnel Stage	Metrics	Technique	PV/Day	Conversion
Aware of web desinger conference	Email sent; email open; social media page visit; ads impression	email; social media; paid search; referal; online display	50000	#VALUE!
Visit Da-shu-ju.com Home Page	External Clicks (email, social media); PV; Visit; Vistor	A/B Testing ad conversion rate;	5000	10%
Visit Buy Ticket Page	Button Click; Click Rate; PV; Visit; Conversion	A/B testing Click Conversion Rate; Promotion/Incencive; Retargeting	1500	30%
Visit Payment Page	Click, PV, Conversion	A/B testing Click Conversion Rate; Promotion/Incencive; Retargeting	1000	67%
Complete Purchase	Conversion; Tickets Sold; Ticket Revenue	A/B testing Click Conversion Rate; Promotion/Incencive; Retargeting	500	50%



Define Metric

	Metric	Definition	Invariant Metric	Evaluation Metric
1	PV of Home Page	# of Pave view of da-shu-ju.com/index.html(home page)	x	
2	Unique Visitors	# of Unique Cookies to view da-shu-ju.com home page	x	
3	"Buy Ticket" Clcik Thru Probablity	# of PV with click on "Buy Ticket"/# of PV of home page		x
4	"Payment" Completion Probability	# of succesful payments/# of PV of "Buy Ticket" page		x
5	Gross Conversion	# of succesful payments/# of PV of home page		x

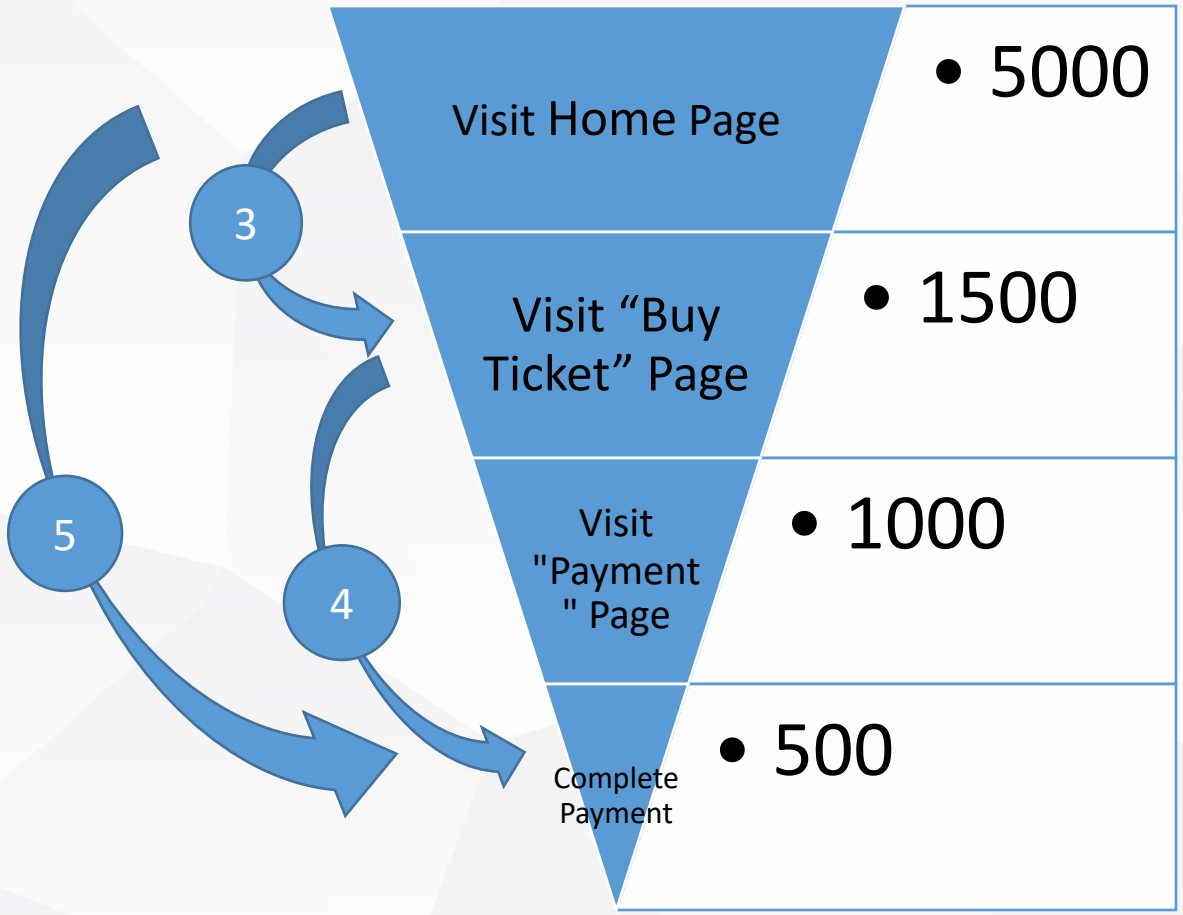


Form Hypothesis

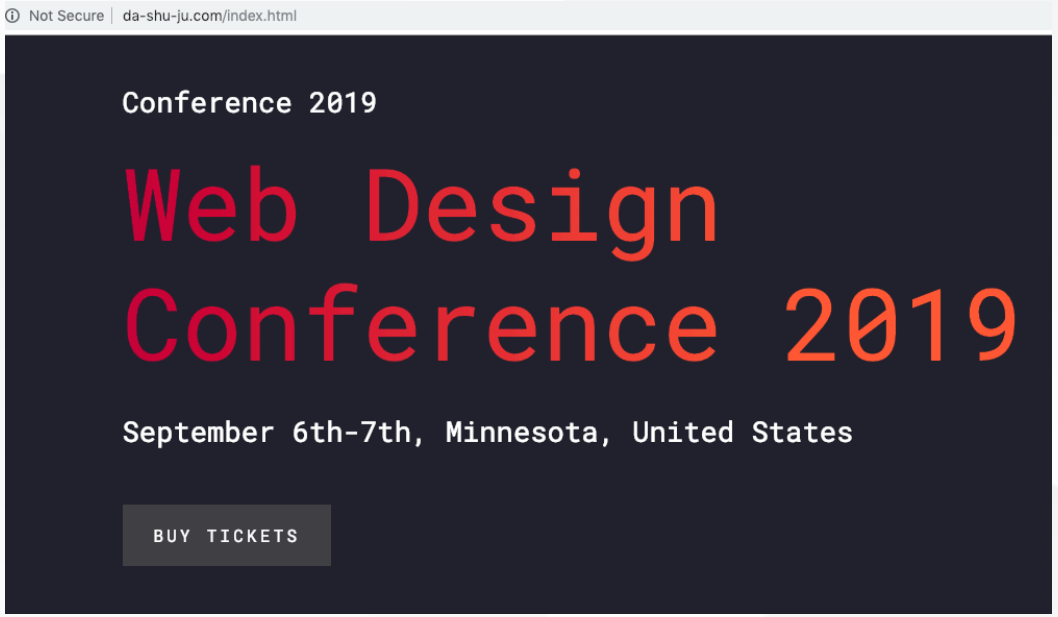
- Change button color to orange
- Change text color
- Increase button size
- Center button
- Change background color
- Change text
- **Change button to “Register”**

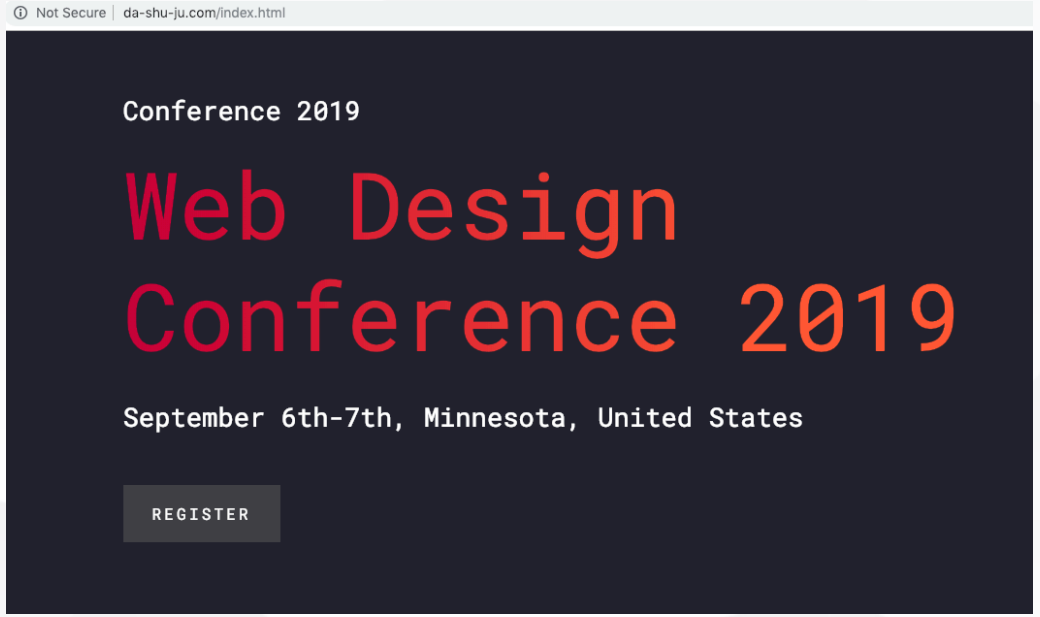
Design Experiment

Hypothesis	Change button to "Register" will increase clicks		
Invariant Metric	PV; Unique Visitor		
Evaluation Metric	Gross Conversion;		
Scope/Target	US; English; Web		
Unit of Division	Cookie		
Sample Size	14313		
Duration	29		
	5. Gross Conversion	3. "Buy Ticket" Clcik Thru Probablity	4. "Payment" Completion Probability
Baseline PV	5000	5000	1500
Baseline Conversion	10%	30%	33%
Practical Significance	0.01	0.05	0.01
Sample Size/Variation	14313	1335	34785
Percent of traffic for testing	20.00%	20.00%	100.00%
Days to run the experiment	28.626	2.67	46.38



Create Variation

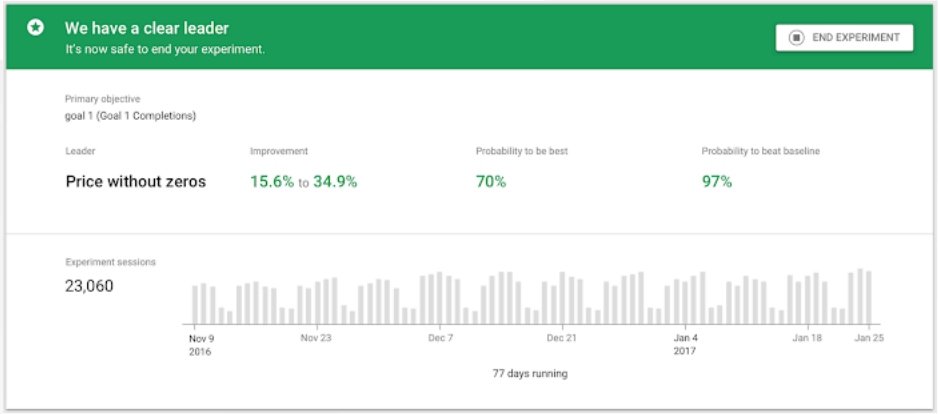




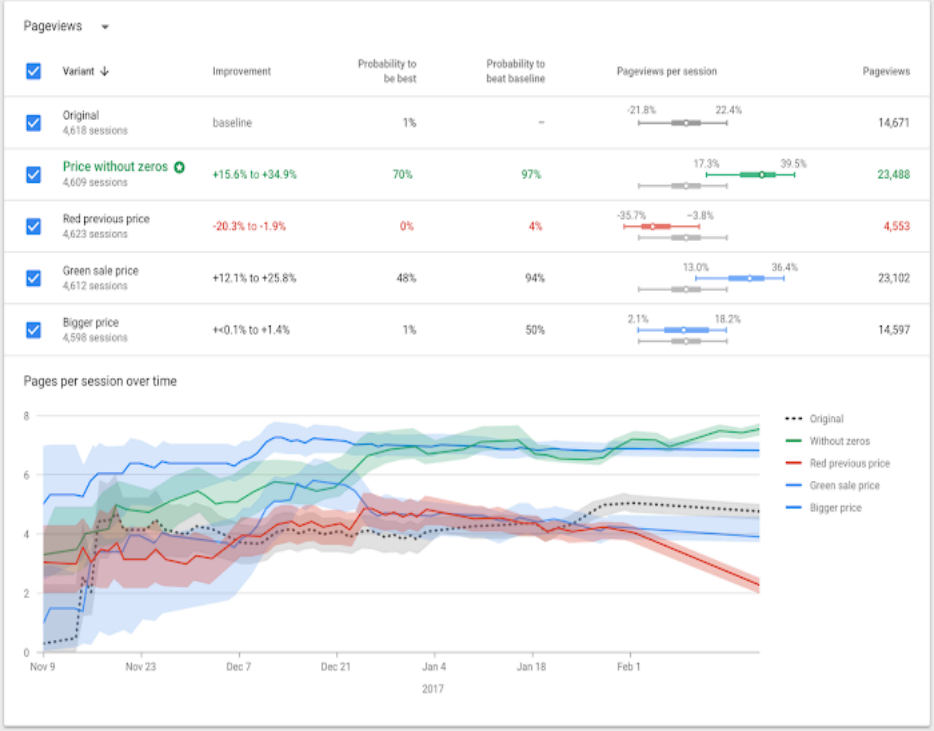
https://optimize.google.com

Run Experiment

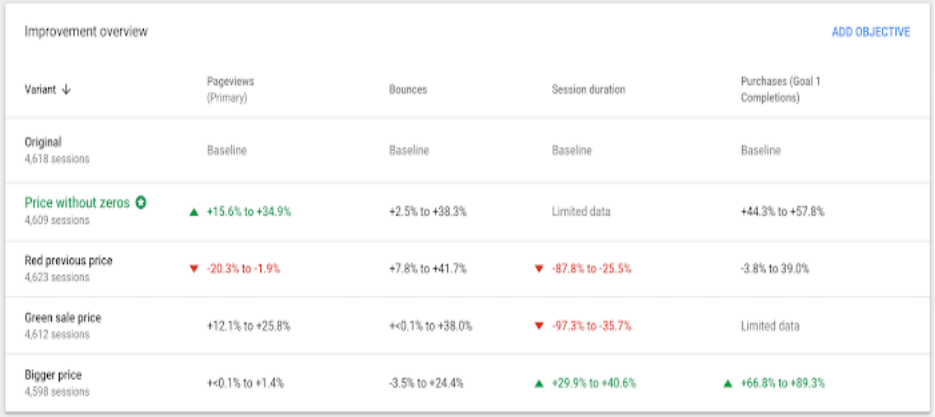
Summary card



Objective detail card



Improvement overview card



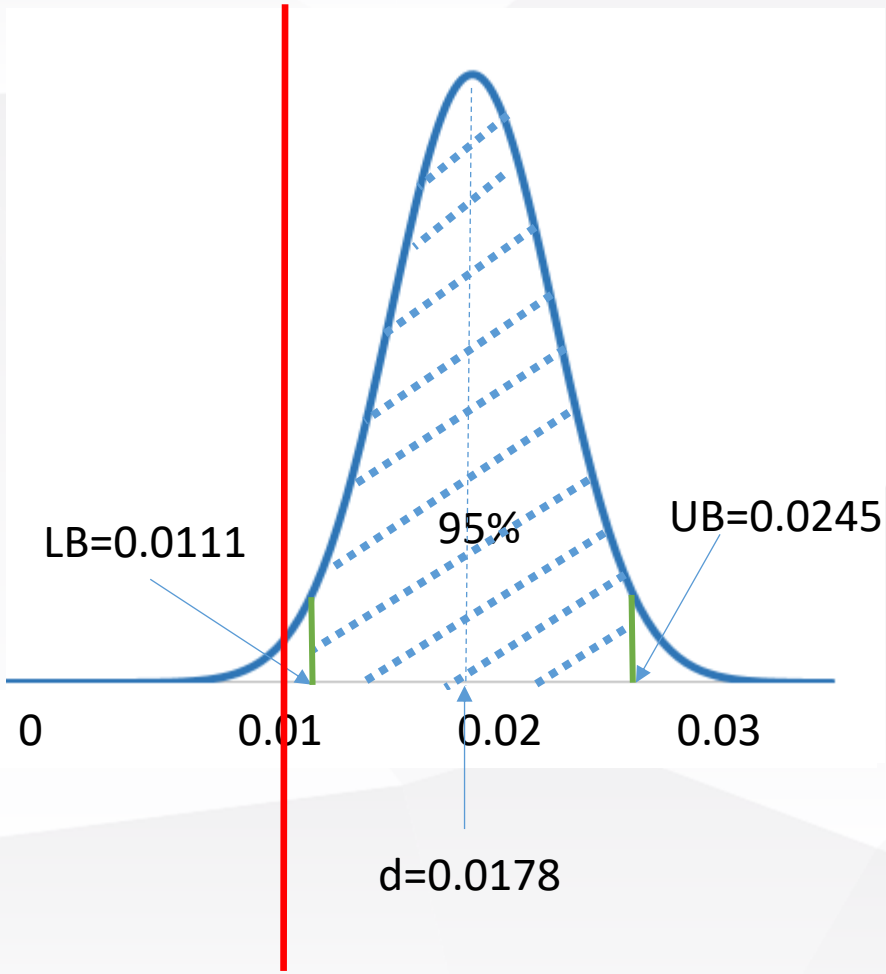
Variant	Pageviews (Primary)	Bounces	Session duration	Purchases (Goal 1 Completions)
Original 4,618 sessions	Baseline	Baseline	Baseline	Baseline
Price without zeros 4,609 sessions	▲ +15.6% to +34.9%	+2.5% to +38.3%	Limited data	+44.3% to +57.8%
Red previous price 4,623 sessions	▼ -20.3% to -1.9%	+7.8% to +41.7%	▼ -87.8% to -25.5%	-3.8% to 39.0%
Green sale price 4,612 sessions	+12.1% to +25.8%	+<0.1% to +38.0%	▼ -97.3% to -35.7%	Limited data
Bigger price 4,598 sessions	+<0.1% to +1.4%	-3.5% to +24.4%	▲ +29.9% to +40.6%	▲ +66.8% to +89.3%

Sample Size Sanity Check

Date	Control - PV	Control - Payment	Experiment - PV	Experiment - Payment
2019/2/1	644	61	643	91
2019/2/2	759	61	774	102
2019/2/3	876	83	873	104
2019/2/4	823	92	822	96
2019/2/5	835	56	816	79
2019/2/6	806	72	792	88
2019/2/7	751	67	757	84
2019/2/8	620	61	639	81
2019/2/9	705	53	703	102
2019/2/10	889	85	875	94
2019/2/11	888	92	879	88
2019/2/12	829	81	811	83
2019/2/13	694	49	681	102
2019/2/14	786	107	784	90
2019/2/15	724	112	722	79
2019/2/16	741	91	740	97
2019/2/17	795	109	805	99
2019/2/18	780	80	783	105
2019/2/19	777	75	772	83
2019/2/20	779	66	776	92
2019/2/21	741	88	726	82
Total	16242	1641	16173	1921

Santity Check:	
p	0.5000
Standard Error	0.0028
Margin of Error	0.0054
Lower Bound CI	0.4946
Upper Bound CI	0.5054
Experiment Size/ To	0.4989

Effective Size Test



Test Result

- Na: 16242, Nb: 16173
- Xa: 1641, Xb: 1921
- Confidence Level: 95%
- Practical Significance: 0.01

Analysis

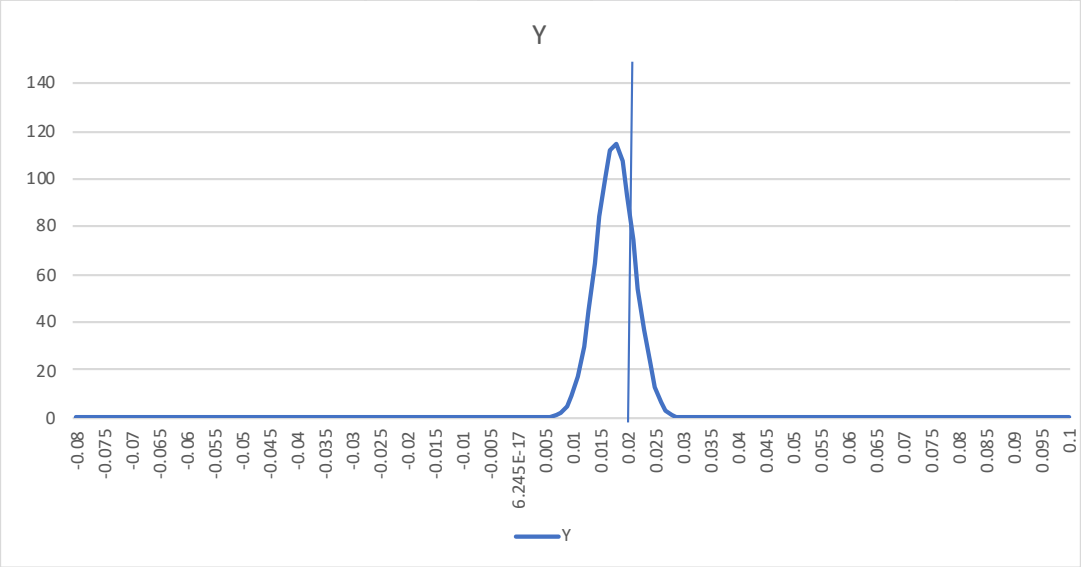
- Pa = 1641/16242 = 0.1010
- Pb = 1921/16173 = 0.1188
- d = 0.1188 – 0.1010 = 0.0178
- Ppool = (1641+1921)/(16242+16173) =0.1099
- SE = $\sqrt{Ppool * (1 - Ppool) * (\frac{1}{Na} + \frac{1}{Nb})}$ = 0.0034
- m = z*SE = 1.96 * 0.0034 = 0.0067
- Confidence Interval= (d-m, d+m) = (0.0111, 0.0245)

Conclusion

- Lower Bound CI > 0
  => Pass Statistical Significance
- Lower Bound CI > Practical Significance (0.0111 > 0.01)
  => Pass Practical Significance

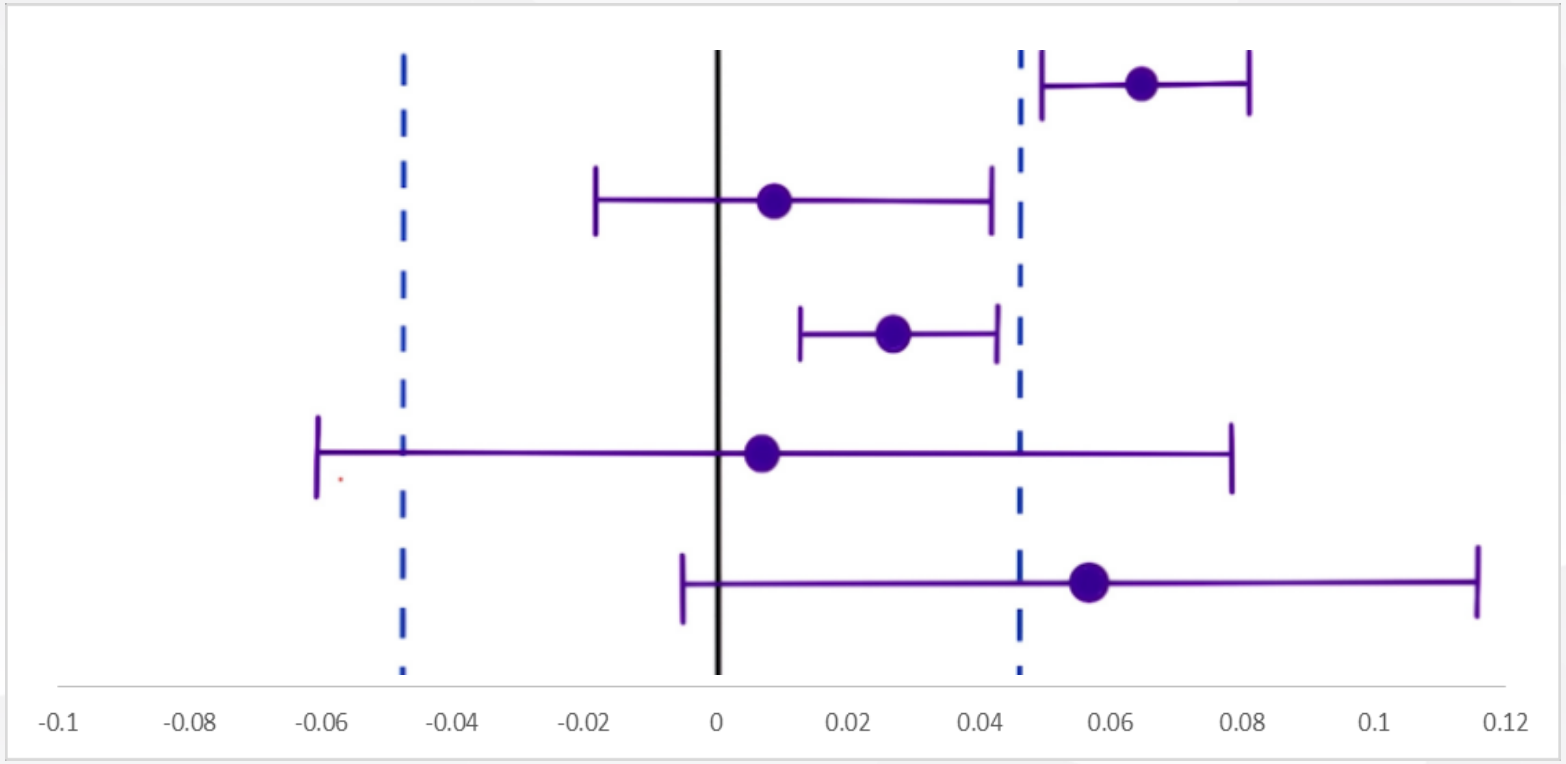
Quiz: A/B Testing Significance Calculator

Confidence Level	95%	Practical Significance	0.02
Group A (Control)		Group B (Variance)	
Size	16242	Size	16173
Conversion	1641	Conversion	1921
Converson Rate	0.101034	Conversion Rate	0.118778
Difference	0.017744		
Pooled Probability	0.109887		
Pooled Standard Error	0.003474		
Lower Bound Confidence Interval	0.010935	Upper Bound Confidence Interval	0.024553
Conclusion			
Need more tests			



Quiz

- Which would you launch given blow test results?



1 Yes

2 No

3 No

4 Need more tests

5 Need more tests

Sanity Check: Sign Test

Date	Control - PV	Control - Payment	Experiment - PV	Experiment - Payment	Experiment > Control
2019/2/1	644	61	643	91	TRUE
2019/2/2	759	61	774	102	TRUE
2019/2/3	876	83	873	104	TRUE
2019/2/4	823	92	822	96	TRUE
2019/2/5	835	56	816	79	TRUE
2019/2/6	806	72	792	88	TRUE
2019/2/7	751	67	757	84	TRUE
2019/2/8	620	61	639	81	TRUE
2019/2/9	705	53	703	102	TRUE
2019/2/10	889	85	875	94	TRUE
2019/2/11	888	92	879	88	FALSE
2019/2/12	829	81	811	83	TRUE
2019/2/13	694	49	681	102	TRUE
2019/2/14	786	107	784	90	FALSE
2019/2/15	724	112	722	79	FALSE
2019/2/16	741	91	740	97	TRUE
2019/2/17	795	109	805	99	FALSE
2019/2/18	780	80	783	105	TRUE
2019/2/19	777	75	772	83	TRUE
2019/2/20	779	66	776	92	TRUE
2019/2/21	741	88	726	82	FALSE
Total	16242	1641	16173	1921	

Sign Test	
# of True	16
# of experim	21
Probability	0.50
P-Value	0.0266
Conclusion	Statistical Significant

[Binomial Sign Test Calculator]

Would you recommendation the launch the variance?

	Metric	Criteria	Observed Value	Conclusion
Sanity Check of Invariant Metric	One Group Sample Size/Total Sample Size	(0.4946,0.5054)	0.4989	Pass
Statistical Significance of Evaluation Metric	Confidence Interval of the difference	>0	(0.0111, 0.0245)	Pass
Practical Significance of Evaluation Metric	Confidence Interval of the difference	>0.1	(0.0111, 0.0245)	Pass
Sign Test of Positive Changes	Probability of days with higher conversion rate (p-value)	<0.05	0.0266	Pass
Final Recommendation	Launch the variance!			

Master A/B Testing Interviews

1. When is A/B testing a good idea? When is it a bad idea?
2. What Is Null Hypothesis/Statistical Significance?
3. What are the main steps in A/B testing?
4. What metrics do you want to improve?
5. Can you think of some variations can be made?
6. What change do you want to test? Why?
7. How many visits/sample size do I need?
8. How long do I need to run A/B testing?
9. Can you calculate the practical significance (given costs)?
10. Will you launch the variation (given the test result)?

Resource

- Udacity A/B Testing Course
  - The most well-known & complete online A/B testing course
- A Summary of Udacity A/B Testing Course
  - Summary of the above course
- The Math Behind A/B Testing with Example Python Code
  - Python implementation of A/B Testing (technical)
- A/B Testing: Introduction & A Complete Guide
  - Business use case & considerations around A/B Testing (non-technical)