一. Sanity Checks

1. propose

- Something may went wrong in the experiment diversion and experiment and control aren't comparable.
- Or setup different filter in the experiment and control
- Or data capture isn't capture events that you want to look for in experiment

2. Two types of check:

(a) population sizing metrics: based on unit of diversion

Check experiment populations and control populations are actually comparable.

(b) Invariant metrics: should not change

Check if they change and different in experiment/control

- Only after all those can we analyses the results

3. Choosing invariants metrics

- same when user have similar behavior not clear cookie more often or view few pages)

	# signed in users	# cookies	# events	CTR on "Start Now"	Time to complete
Changes order of courses in course list Unit of diversion: user-id ∵ it's unit of diversion	☑ random	☑ not directly but should	☑ randomized be split easily	☑ happens before course list	X could be affected
Changes infrastructure to reduce load time Unit of diversion: event	☑ "larger" than unit of diversion	☑	☑ random ∵ it's unit of diversion	☑ happens before viewing videos	X can't be tracked

- one user or one cookie could correspond to multiple events. ∵ event random, they should be same
- may be those exist learning effect, but could not capture when divert it by event
- choose list may affect user enroll in different courses.
- Users can both be assigned to experiment and control multiple times

Choosing invariant metrics

Experiment: Change location of ~~sign-in button~~ to appear on every page Unit of diversion: Cookie

Which metrics would make good invariants? will not change because of experiment

- [x] # events — This, #cookies, and #users all good
- [ ] CTR on "Start Now" — Adding sign-in button to home page could affect this
- [ ] Probability of enrolling — Users often enroll after signing in
- [ ] Sign-in rate — This is what we're trying to change!
- [x] Video load time — No backend changes

4. Checking invariants

Checking invariants

Run experiment for 2 weeks.
Unit of diversion: cookie

Total control: 64,454
Total experiment: 61,818

Is that difference within expectations?

Week1:

Day	#cookies control	#cookies experiment
Mon	5077	4877
Tue	5495	4729
Wed	5294	5063
Thu	5446	5035
Fri	5126	5010
Sat	3382	3193
Sun	2891	3226

Week2:

Day	#cookies control	#cookies experiment
Mon	5029	5092
Tue	5166	5048
Wed	4902	4985
Thu	4923	4805
Fri	4816	4741
Sat	3411	2939
Sun	3496	3075

(1) Firstly, compute the total number for each group, see if overall division looks even

How would you figure out whether this difference is within expectations?

Given: Each cookie is randomly assigned to the control or experiment group with probability 0.5

Like flipping a fair coin: heads → control

(Sample size large enough to consider as normal)

① Compute standard deviation of binomial with probability 0.5 of success $SD = \sqrt{\frac{0.5*0.5}{64454+61818}} = 0.0014$

② Multiply by z-score to get margin of error $m = SD*1.96 = 0.0027$

③ Compute CI around 0.5 [0.4973, 0.5027] fraction = 0.5104, out of interval

④ Check whether observed fraction is within this interval

(2) When not look even, need to look at day by day data

{ check if any particular day stands out as causing the problem)
  check if there is an overall pattern

Checking invariants

Run experiment for 2 weeks.
Unit of diversion: cookie

Total control: 64,454
Total experiment: 61,818
More cookies in control:

Week1:

Day	# cookies control	# cookies experiment	$\hat{p}$
Mon	5077	4877	0.510
Tue	5495	4729	0.537
Wed	5294	5063	0.511
Thu	5446	5035	0.520
Fri	5126	5010	0.506
Sat	3382	3193	0.514
Sun	2891	3226	0.473

Week2:

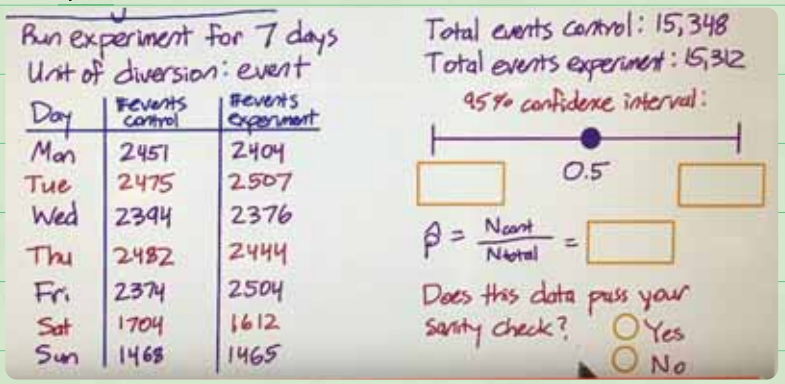
Day	# cookies control	# cookies experiment	$\hat{p}$
Mon	5029	5092	0.497
Tue	5166	5048	0.506
Wed	4902	4985	0.496
Thu	4923	4805	0.506
Fri	[illegible]	4741	0.504
Sat	[illegible]	[illegible]	0.537
Sun	3446	[illegible]	0.532

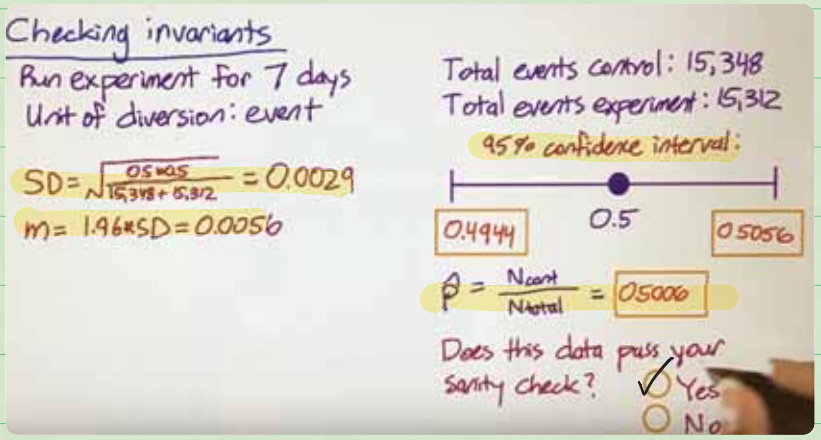
∵ many days are higher and no specific day for extremely high
∴ It's an overall problem

(3) what to do.

- Talk to engineers
- Try slicing to see if one particular slice is weird.
- Check age of cookies - does one group have more new cookies

• Example for sanity check





5. What to do if sanity check fail

• Don't proceed, go straight to analyzing why sanity check fail.

(1) Figure out what went wrong

① Technically: Is there something wrong with the experiment infrastructure?

↑ Is something wrong with experiment diversion?

Experiment set up, handle with engineer

② Retrospective analysis: Try and recreate experiment diversion from data capture to see if there is something endemic

③ Use pre and post period

- If also see the problem in pre-period, that points to a problem with experiment infrastructure, the set up, something alone those lines.
- If just in experiment. that points to a problem with the experiment itself, eg: data capture. ~

(2) Common problem

① Data capture, especially for a new experience that the user is undergoing. (default capture; change triggers very rarely and just capture it correctly in experiment, but don't capture quickly in control)

② Experiment set up:
Filter applied to a group of users only.

③ Infrastrue, experiment system

(3) key points to remember:

① All the comparison are approximation; just get approximately the same

② Maybe it's due to a learning effect

- If it's true, there are not much change at beginning and it's going to be increasing over time.
- If there is a big change at beginning, probably not a learning effect

二. Single metric:

1. Example 1

Note: ① CTR → possion distribution, need to calculate SE empirically

② $SE \sim \frac{1}{\sqrt{N}}$ is used to when experiment and control have same size

$SE \sim \sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$ should be used for different sample size

(when $N_1 = N_2 = N$, is $\sqrt{\frac{1}{N}+\frac{1}{N}}$)

① Analysis with a single metric (parametric way)

Experiment: Change color and placement of "Start Now" button
Metric: Click-through-rate
Unit of diversion: cookie

$d_{min} = 0.01$
$\alpha = 0.05$ $\beta = 0.2$

	control clicks	control pageviews	experiment clicks	experiment pageviews
Day 1	51	1242	115	1305
Day 2	39	853	73	835
Day 3	64	1129	91	1133
Day 4	43	873	60	871
Day 5	55	1197	78	1134
Day 6	44	1023	72	1015
Day 7	56	1003	76	977
Total	352	7370	565	7270

Sanity check: pass

Empirical SE: 0.0035 w/ 10,000 pageviews per group

$SE \sim \frac{\sigma}{\sqrt{N}} \sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$

$\frac{0.0035}{\sqrt{\frac{1}{10000}+\frac{1}{10000}}} = \frac{SE}{\sqrt{\frac{1}{7370}+\frac{1}{7270}}}$

$SE = 0.0041$

Calculate SE when sample size change

• standard derivation of original data

① calculate CTR each day

② get an std (σ) for all the CTR

③ $SE = \frac{\sigma}{\sqrt{N}}$, N represent for total number

$\hat{d}$: use number in both group to calculate

$X_{cont} = 352$, $N_{cont} = 7370$; $X_{exp} = 565$, $N_{exp} = 7270$

$\hat{d} = \hat{r}_{exp} - \hat{r}_{cont} = \frac{565}{7270} - \frac{352}{7370} = 0.0300$

$m = 0.0041 * 1.96 = 0.0080$ ∴ CI = [0.022, 0.0380]

Recommend since CI close not include practical significance boundary

② Double check by non-parametric way

	control clicks (CTR)	control pageviews	experiment clicks (CTR)	experiment pageviews
Day 1	51 (.039)	1242	115 (.088)	1305
Day 2	39 (.046)	853	73 (.087)	835
Day 3	64 (.057)	1129	91 (.080)	1133
Day 4	43 (.049)	873	60 (.069)	871
Day 5	55 (.046)	1197	78 (.069)	1134
Day 6	44 (.043)	1023	72 (.071)	1015
Day 7	56 (.056)	1003	76 (.078)	977
Total	352 (.048)	7370	565 (.078)	7270

Sanity check: pass

days: 7

days with positive change: 7

If no difference, 50% chance of positive change on each day

Note: We cannot assume normal since we just have 7 days

- Use sign test calculator:

Get two-tail P value vs 0.0156 ∵ $P<0.05$ ∴ unlikely happened by chance.

Meaning of P-value: probability of observing a result at least this extreme by chance.

2. Example 2: (Different in parametric and non-parametric results)

Analysis with a single metric

Metric: click-through-rate $d_{min}=0.01$ $\alpha=0.05$

Additional information given: Empirical SE: 0.0062 with 5000 pageviews in each group

Week 1

	Control CTR	Control #pageviews	Experiment CTR	Experiment #pageviews
Mon	0.097	2029	0.091	1971
Tue	0.100	1991	0.104	2009
Wed	0.103	1951	0.100	2049
Thu	0.109	1985	0.087	2015
Fri	0.107	1973	0.094	2027
Sat	0.092	2021	0.147	1979
Sun	0.110	2041	0.142	[illegible]

Week 2

	Control CTR	Control #pageviews	Experiment CTR	Experiment #pageviews
Mon	0.094	1980	0.107	2020
Tue	0.105	1951	0.110	2049
Wed	0.106	1988	0.103	2012
Thu	0.097	1977	0.101	2023
Fri	0.097	2019	0.101	1981
Sat	0.110	2035	0.151	1965
Sun	0.096	2007	0.150	1993

① Analysis with a single metric (parametric way to calculate effective size)

Metric: click-through-rate $d_{min}=0.01$ $\alpha=0.05$

Empirical SE: 0.0062 with 5000 pageviews in each group

Control pageviews: 27,948 Control CTR: 0.1016

Experiment pageviews: 28,052 Experiment CTR: 0.1132

$\hat{d}=0.1132-0.1016=0.0116$

$$\frac{SE}{\sqrt{\frac{1}{27,948}+\frac{1}{28,052}}}=\frac{0.0062}{\sqrt{\frac{1}{5000}+\frac{1}{5000}}} \qquad SE=0.0026$$

$m=0.0026*1.96=0.0051$ Confidence Interval: 0.0065 to 0.0167

∵ CI don't include 0, it's statistically significant

∴ Unlikely there is no real difference

But CI include d_min (practical significance boundary), so we can't be confident at 95% level that the size of this effect is something we care about.

② non-parametric way: sign test

Analysis with a single metric

Metric: click-through-rate $d_{min}=0.01$ $\alpha=0.05$

Days where CTR is higher in experiment: 9/14

	Week 1 Control CTR	Control #pageviews	Experiment CTR	Experiment #pageviews		Week 2 Control CTR	Control #pageviews	Experiment CTR	Experiment #pageviews
Mon	0.097	2029	0.091	1971	✓Mon	0.094	1980	0.107	2020
✓Tue	0.100	1991	0.104	2009	✓Tue	0.105	1951	0.110	2049
Wed	0.103	1951	0.100	2049	Wed	0.106	1988	0.103	20[illegible]
Thu	0.109	1985	0.087	2015	✓Thu	0.097	1977	0.101	20[illegible]
Fri	0.107	1973	0.094	2027	✓Fri	0.097	2019	0.101	19[illegible]
✓Sat	0.092	2021	0.147	1979	✓Sat	0.110	2035	0.151	1985
✓Sun	0.110	2041	0.142	1959	✓Sun	0.096	2007	0.150	1993

Use calculator, get two tailed P value is 0.4240

$p > 0.05$ ∴ Not significant

Analysis with a single metric

Metric: click-through-rate $d_{min}=0.01$ $\alpha=0.05$

Empirical SE: 0.0062 with 5000 pageviews in each group

Effect size

0.0065 — 0.0167

Statistically significant?

(✓) Yes ( ) No

Sign test

p-value: 0.4240

Statistically significant?

( ) Yes (✓) No

(price paid for don't make any hypothesis)

Sign test has lower power, but let's dig deeper

∴ The difference in result is not necessary a red flag

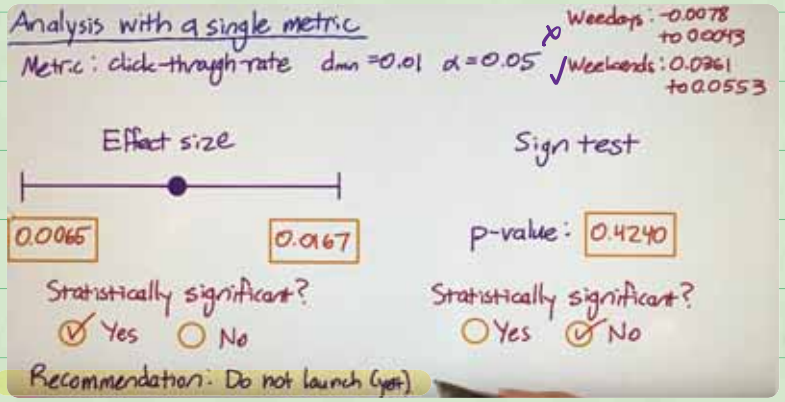
Back to check CTR in different days, notice that it's much higher on weekends

∴ The change may not have a significant impact on weekdays

but a significant impact on weekends.

Then, calculate CI seperatly { weekdays: $[-0.0078, 0.0043]$ × significant; weekends: $[0.0361, 0.0553]$ ✓ significant

The sign test also gives same result.



- Need to dig deeper about into why the change didn't affect weekday visitors
  Once understand that, may know how to iterate on the change to help it affect more users
- If not, need to talk to decision makers.

3. Notches

① Take a critical look about how your feature functions:
Maybe different on different subgroups (eg: different plat form)

② Statistical reason: Simpson's paradox

Def: Within each subgroup, the results are stable but when you aggregate them all together, it's the mix of subgroups that drives your results.

③ Simpson's paradox: example

- UCB Admission Example 在同department 都对女性有，但合起来而错了.

	Men applied	Women applied	Men accepted	Women accepted
Department A	825	108	512 (62%)	89 (82%)
Department B	417	375	137 (33%)	132 (35%)
Total	1242	483	649 (52%)	221 (46%)

Because more women applied to department with low acceptance rate.

即为：更倾向于申请的群组中占比较大

- A-B testing example

	Ncont	Xcont (CTR)	Nexp	Xexp (CTR)
New Users	150,000	30,000 (.02)	75,000	18,750 (.025)
Experienced Users	100,000	1,000 (.001)	175,000	3,500 (.02)
Total	250,000	31,000 (.0124)	250,000	22,250 (.089)

Goal: click-through-rate is higher in experiment group for both new and experienced users, but overall click-through-rate is lower in the experiment group

Note: number of page view should be evenly split (for sanity check)

- sanity check to make sure pageview is same for experiment group and control group.
- check the breakdown across different slices should also be a sanity check.

Why the pageview is different?

- Something went wrong with the set up
- change affects new users and experienced users differently

Eg: Diverting based on user_id; change makes new users generate less pageviews.

Recommendation: good change since both CTRs increase,
but need to dig deeper about why pageviews are different

三. Multiple Evaluations.

1. If have more metrics, more likly to see significance by chance.
but it's not repeatable (do same experiment, will have different result)

- Solution: Use Multiple comparison to adjust significance level
- When to use? Do automatic detection of differences
- When do EDA, can reanalyze your data and make sure same metric won't show every time and see if the differences are repeatable

2. Example 1: Tracking multiple metrics (Bonferroni)

Tracking multiple metrics

Experiment: Prompt students to contact coach more frequently

Metrics:

- Probability that student signs up for coaching
- How early students sign up for coaching
- Average price paid per student

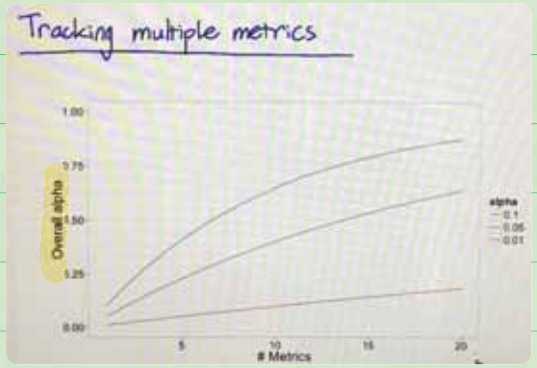
If Audacity tracks all three metrics and does three separate significance tests ($\alpha = 0.05$), what is the probability at least one metric will show a significant difference if there is no true difference?

$P(FP = 0) = 0.95 * 0.95 * 0.95 = 0.875$ (Assuming independent)

$P(FP \geq 1) = 1 - 0.857 = 0.143$ (is overestimate based on assumption)

$\alpha_{overall} = 1 - (1 - \alpha_{ind})^n$

Tracking multiple metrics



Tracking multiple metrics

Problem: Probability of any false positive increases as you increase number of metrics

Solution: Use higher confidence level for each metric

Method 1: Assume independence

$\alpha_{overall} = 1 - (1 - \alpha_{individual})^n$

Method 2: Bonferroni correction, $\alpha_{ind} = \frac{\alpha_{overall}}{n}$

- simple
- no assumptions
- conservative — guaranteed to give $\alpha_{overall}$ at least as small as specified

eg: $\alpha_{overall} = 0.05$, $n = 3$ $\Rightarrow \alpha_{ind} = 0.0167$

- Problem of Bonferroni:

When tracking metrics that are correlated and tend to move at the same time, this method is too conservative, which will lead to company launching fewer experiments than they would like.

3. Example 2 (Bonferroni too conservative)

① Tracking multiple metrics

Bonferroni: $\alpha_{indiv} = \alpha_{overall}/n$

Experiment: Update description on course list

$z^* = 1.96$ Statistically significant? $z^* = 2.5$ ← $z\ 1-\frac{\alpha}{2}$

metrics	$\hat{d}$	SE	$\alpha_{indiv} = 0.05$	Bonferroni $\alpha_{overall} = 0.05$
prob of clicking through to course overview	0.03	0.013	☑ .02548	☐ .0325
avg time spent reading course overview page	-0.5 s	0.21	☑ .4116	☐ .5250
prob of enrolling	0.01	0.0045	☑ .0088	☐ .0113
avg time in classroom during first week	10 min	6.85	☐ 13.43	☐ 17.13

Is Bonferroni overly conservative here? ☑ Yes ○ No

② Ways to handle the conservative

Tracking multiple metrics

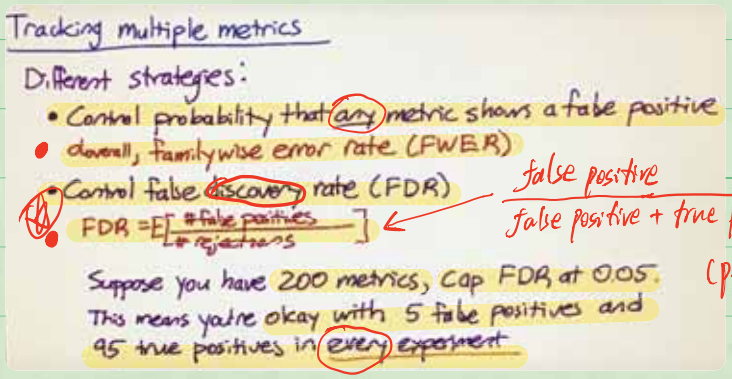
Experiment: Update description on course list
3 out of 4 metrics had significant difference at $\alpha = 0.05$, but none were significant using Bonferroni correction

Recommendation:
Rigorous answer: Use a more sophisticated method
In practice: Judgment call, possibly based on business strategy

Communicate to decision makers

4. Other techniques (False Discovery rate)

If want to detect significant changes across a large number of metrics, then capping the false discovery rate instead of the familywise error rate can be more lenient.



5. Analyzing Multiple Metrics

① Related metrics tend to move same direction:

CTR, CTP

② Composite metric:

• RPM (revenue per thousand queries) is compost of CTR and cost per click

Analyze why RPM more based on its construction

③ But sometimes different metrics go opposite, that is why we want an OEC (overall evaluation criteria)

eg: More CTR but spend less time on news website

• Question about OEC:

(a) how to find a good one.

(how to balance stade time and click)

(b) Don't absorb why they move different direction
but helpful with this balancing long term investment,
like return visits to the site, with short term day to day
metrics like increased clicks.

④ How to come up an overall metric

- Based on analyst

(a) start with business analyst and the goal

(b) After getting a few candidate, need to run a whole bunch of different experiments
and validate how they steer you (Are they in right direction)

(c) Downside of it:
Don't plan so much around what the company thinks should happen
that steer yourself in a way that you hide other changes that other
changes that you weren't expecting.

- Based on test results and stakeholders

Give experiment results to decision makers, didn't tell them what the
experiment is testing and actual launch decision
Let them discuss and than change weight of OEC

⑤ Goal of OEC
Tell everyone what we care about, but always need to look out individual
metrics before make decisions

6. Drawing Conclusions

① { Do you understand the change ?
  Do you want to launch the change ?

② If some metrics are significant but others are not ?

Need intuition and other experiments results

- For small changes, a change in one metrics but no change all the other of time.
  But for big changes, probably indicate something is wrong.

③ If positive for one slice of users but negative for other slice

Need to understand why

④ How do you decide whether to lauch a change or not

(a) Do I have statistically and practically significant results in order to justify changes?

(b) Do I understand what the change actually done for user experience ?

(c) Is that worth it ?

- End goal:

Making that recommendation that shows your judgment.

6. Gotchas: Disappearing launch effect

1. The effect may actually flatten out as you ramp up the change.

Seasonality effect: use hold back to handle it.

(cacert-driven) A set of users that don't get the change and continue comparing their behavior to the control

∴ Can see the reverse of impact to your experiment

Track that over time until you confident that your results are repeatable

2. Novelty effect / change of aversion

- As users discover or adopt or change the adoption of your change, then their behavior can change and therefore the measured effects can change
- Use Cohort Analysis to handle that

3. Have advertisment that have budgets,

if don't control for budgets properly, the effect can change as you ramp up.

(4) How to handle it during experiment

Use pre and post periods, combined with a cohort analysis, to try to understand how users basically adopting to the change.

2. Lesson learned

1. Check the experiment is set up properly:

   look at invariants, check the experiment metrics are actually looking same.

2. Aren't just looking for statistical significance!

   eg: ↑ for 30% but neutral for everyone else

   ↑ for 70% but ↓ for 30%

   - Do you want to launch or try to make it better?

3. Take into account of the business analysis

   Judgement call based on user experiences and business

   - What's the engineering cost of maintaining the change?
   - Are there customer support or sales issues?
   - Overall, what's the opportunity cost if you actually launch the change relative to the reward from the change and not launching the change

3. Conclusion

1. Fundamentals of handling the results of your experiment.
2. Coaches for single and multiple metrics
3. Suggestion for converting significant results into a real business case for experiment.