Analyzing results

Sanity Check: check whether there’re things going wrong by looking at invariant metrics

Examples of things can go wrong:

- Unit of diversion – experiment and control should be comparable
- Set up filters consistently between experiment and control
- Is data capture set up accurately capturing the events you are looking for?

<u>Use invariants to do sanity check – two types:</u>

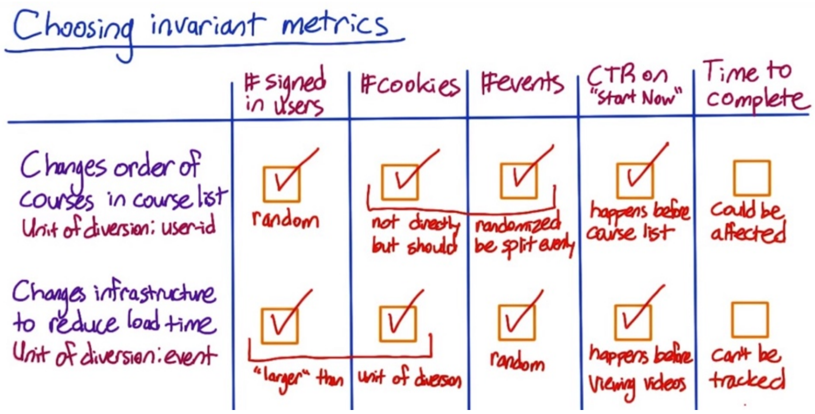
(1) population sizing metrics based on unit of diversion

Experiment population and control population should be comparable.

(2) other invariants – the metrics that shouldn’t change in your experiment

Should test if these metrics change or not.

Should choose the invariant metrics based on the feature you are changing and where it falls under the overall process. E.g. if the feature affects the steps from #4, then the metrics associated with steps before #4 can be used as invariant metrics.



Choosing invariant metrics

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

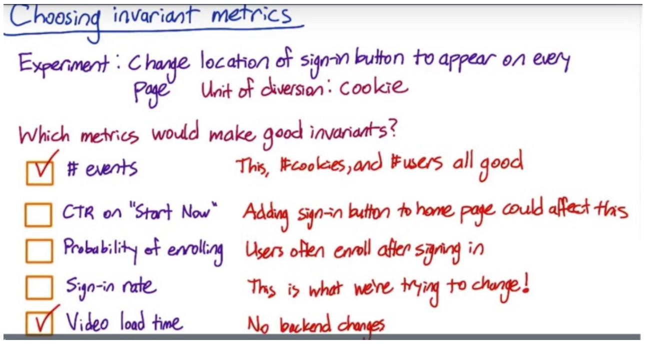
Case #1.

Since user id is used as unit of diversion, which is randomly assigned between 2 groups. Cookies and events might not be exactly the same between control and experiment, but should not vary too much unless users visit the pages significantly differently between the two group. CTR happens before course list. Time to complete might be affected as students might start with easier course based on the new order. Maybe putting easier courses first causes more users to start with easier courses, and then they finish them faster.

Case #2.

Signed in users and cookies are both larger than the unit of diversion in the sense that one user or one cookie could correspond to multiple events. So, since the events are being randomly assigned, the number of signed in users and cookies shouldn't be different between the two groups either.

The time to get through a class can't be tracked if you're using event-based diversion. Since by the time the user gets through a course, they could have been assigned to both the experiment and the control group multiple times. Even if you could track this, it wouldn't be a good invariant, since load time could affect how long it takes to complete a class.



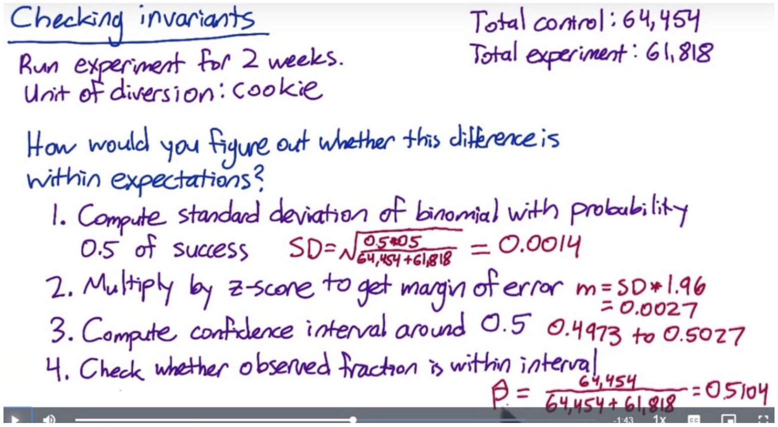
Choosing invariant metrics

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

Which metrics would make good invariants?

☑	# 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!
☑	Video load time	No backend changes

Cookies are being explicitly randomized over. User IDs are typically larger than cookies, in the sense that one user ID can correspond to multiple cookies. So user IDs should be evenly split as well. And it's more likely that the events could end up unevenly split, but it's not something you're expecting. And it would be good to catch that if it does happen.



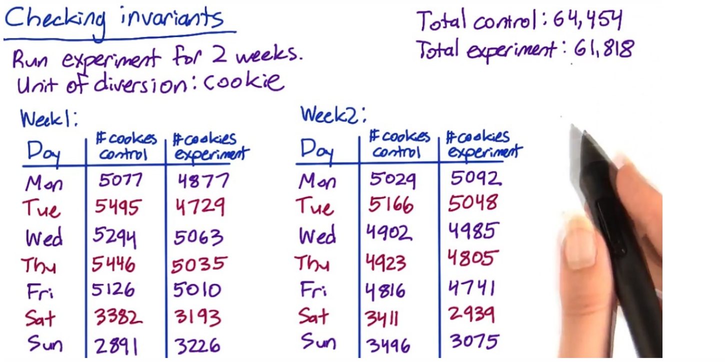
Checking invariants

Run experiment for 2 weeks.
Unit of diversion: cookie

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

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

1. Compute standard deviation of binomial with probability 0.5 of success $SD = \sqrt{\frac{0.5*0.5}{64,454+61,818}} = 0.0014$
2. Multiply by z-score to get margin of error $m = SD * 1.96 = 0.0027$
3. Compute confidence interval around 0.5 0.4973 to 0.5027
4. Check whether observed fraction is within interval $\hat{p} = \frac{64,454}{64,454+61,818} = 0.5104$



Checking invariants

Run experiment for 2 weeks.
Unit of diversion: cookie

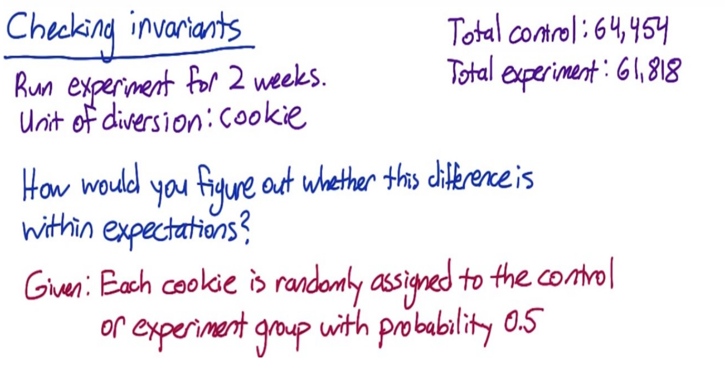
Total control: 64,454
Total experiment: 61,818

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



Checking invariants

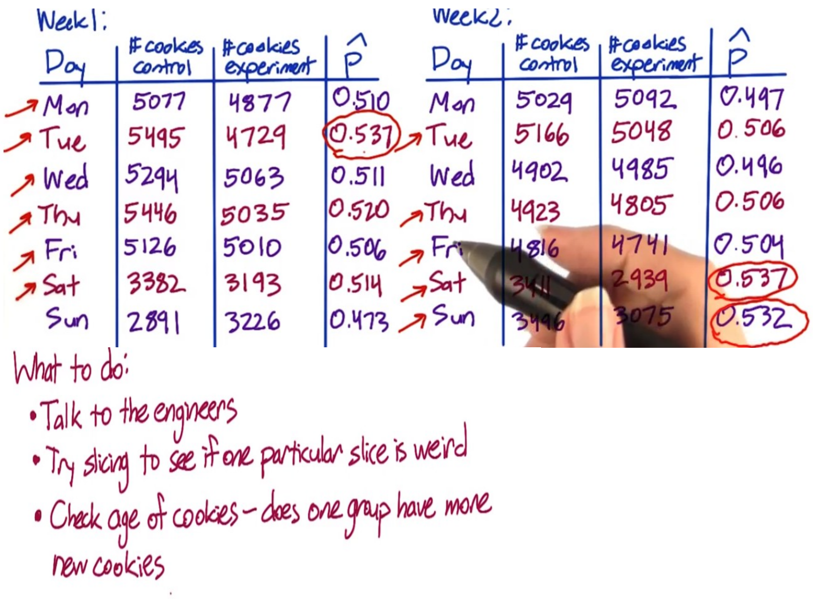
Run experiment for 2 weeks.
Unit of diversion: cookie

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

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

In step 1, the "standard deviation" is the standard deviation of the sampling distribution for the proportion, or standard error. The abbreviation SE should be used in computations instead of SD. Use p=0.5 to calculate the confidence interval. The observed fraction of control group is greater than the upper bound of CI, so there is something wrong with the setup. Do day-by-day analysis:



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	4816	4741	0.504
Sat	3411	2939	0.537
Sun	3496	3075	0.532

What to do:

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

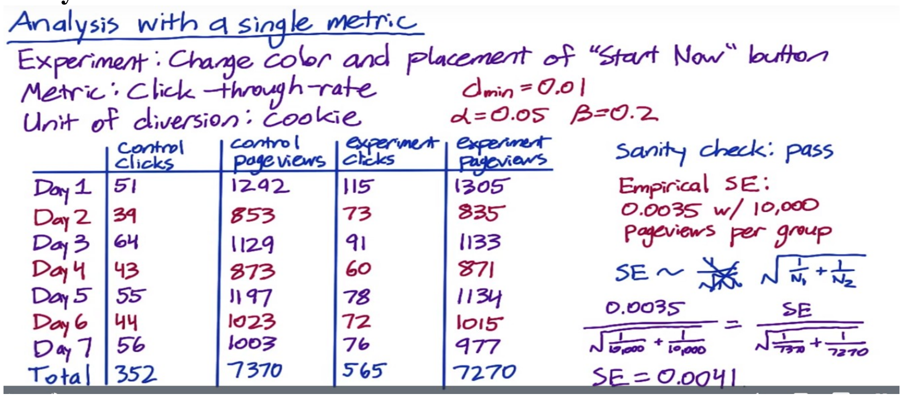
What to do if you find issues during the sanity check
(1) Issues to check with the engineering team:

- Experiment infrastructure
- Unit of diversion

(2) Retrospective Analysis
Recreate the experiment diversion from the data capture to understand if there is something endemic to what you are trying to do that might cause the situation.
(3) Pre and Post period
If observe changes for invariant metrics on post period, check if similar changes exist on pre period. If so, there could be problems with the experiment infrastructure, setup, etc. If the changes is only observed on the post period, it means the issue is associated with the experiment itself such as data capture.

The most common thing – data capture. Maybe the changes trigger rarely, and you capture it correctly under the experiment but not the control.
Experiment setup – didn't set up filter correctly between control and experiment.
More rarely could be system issue such as cookie reset (need to dig deeper and find out with engineering team)
Learning effect may take time. If the issues are observed at the beginning of the experiment, might not be learning effect.

Analyze the results



Analysis with a single metric

Experiment: Change color and placement of "Start Now" button
Metric: Click-through-rate $d_{min} = 0.01$
Unit of diversion: Cookie $\alpha = 0.05$ $\beta = 0.2$

	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

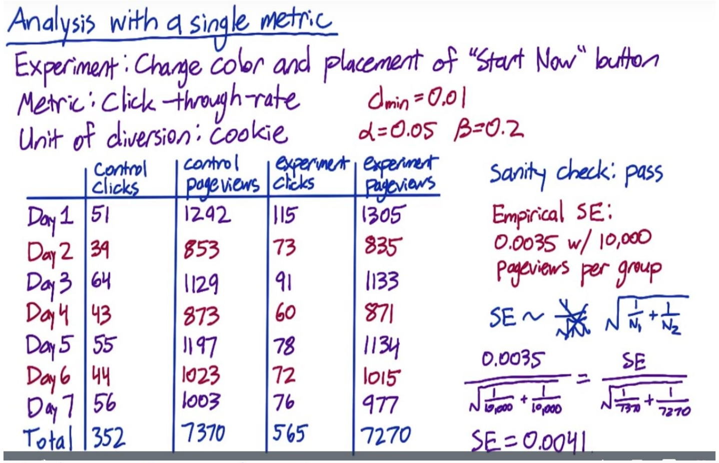
Sanity check: pass

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

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

$$\frac{0.0035}{\sqrt{\frac{1}{10{,}000}+\frac{1}{10{,}000}}} = \frac{SE}{\sqrt{\frac{1}{7370}+\frac{1}{7270}}}$$

$SE = 0.0041$



Analysis with a single metric

Experiment: Change color and placement of "Start Now" button

Metric: Click-through-rate $d_{min} = 0.01$

Unit of diversion: cookie $\alpha = 0.05$ $\beta = 0.2$

	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

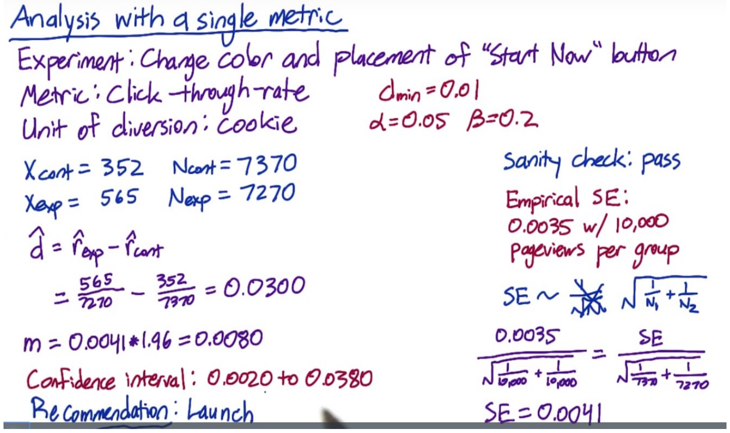
Sanity check: pass

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

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

$$\frac{0.0035}{\sqrt{\frac{1}{10,000}+\frac{1}{10,000}}} = \frac{SE}{\sqrt{\frac{1}{7370}+\frac{1}{7270}}}$$

$SE = 0.0041$



Analysis with a single metric

Experiment: Change color and placement of "Start Now" button

Metric: Click-through-rate $d_{min} = 0.01$

Unit of diversion: cookie $\alpha = 0.05$ $\beta = 0.2$

$X_{cont} = 352$ $N_{cont} = 7370$

$X_{exp} = 565$ $N_{exp} = 7270$

$\hat{d} = \hat{r}_{exp} - \hat{r}_{cont}$

$= \frac{565}{7270} - \frac{352}{7370} = 0.0300$

$m = 0.0041 * 1.96 = 0.0080$

Confidence interval: 0.0020 to 0.0380

Recommendation: Launch

Sanity check: pass

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

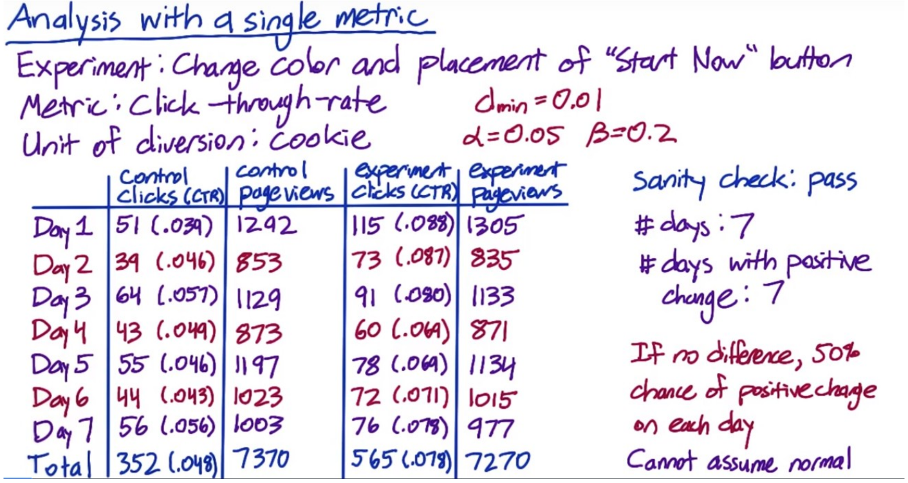
$SE \sim \sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$

$$\frac{0.0035}{\sqrt{\frac{1}{10,000}+\frac{1}{10,000}}} = \frac{SE}{\sqrt{\frac{1}{7370}+\frac{1}{7270}}}$$

$SE = 0.0041$

Unlike click-through probability, click-through rate more follows Poisson distribution and the analytical variance is harder to estimate compared to Binomial. Need to analyze it empirically. Typo: 0.022 – 0.038. Practical significance is out of the CI, which is 0.01, meaning the difference is significant to be captured, and the change should be launched.

Sign Test:
https://en.wikipedia.org/wiki/Sign_test



Analysis with a single metric

Experiment: Change color and placement of "Start Now" button

Metric: Click-through-rate $d_{min} = 0.01$

Unit of diversion: cookie $\alpha = 0.05$ $\beta = 0.2$

	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

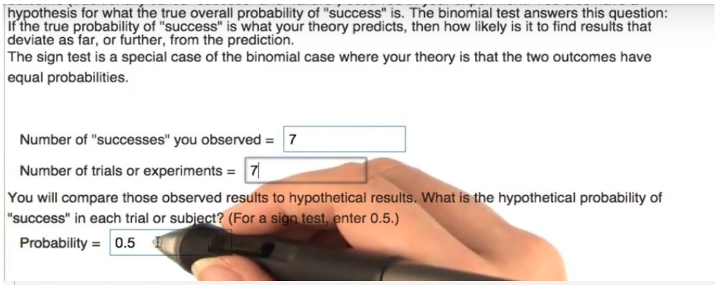
Sanity check: pass

days: 7

days with positive change: 7

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

Cannot assume normal



hypothesis for what the true overall probability of "success" is. The binomial test answers this question: If the true probability of "success" is what your theory predicts, then how likely is it to find results that deviate as far, or further, from the prediction.

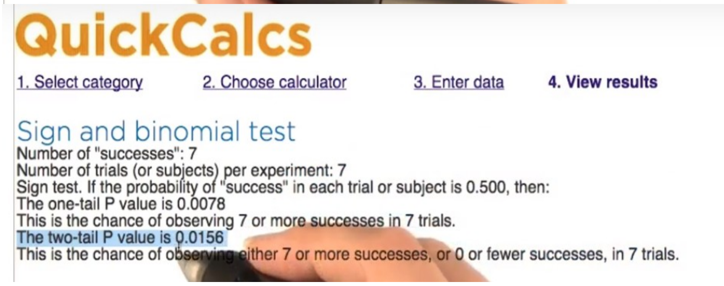
The sign test is a special case of the binomial case where your theory is that the two outcomes have equal probabilities.

Number of "successes" you observed = 7

Number of trials or experiments = 7

You will compare those observed results to hypothetical results. What is the hypothetical probability of "success" in each trial or subject? (For a sign test, enter 0.5.)

Probability = 0.5



QuickCalcs

Sign and binomial test

Number of "successes": 7

Number of trials (or subjects) per experiment: 7

Sign test. If the probability of "success" in each trial or subject is 0.500, then:

The one-tail P value is 0.0078

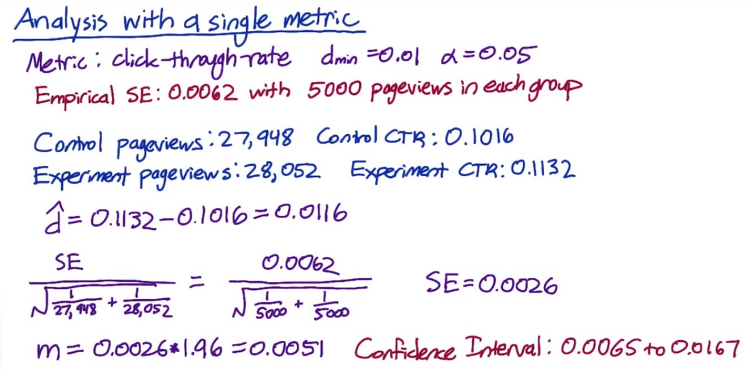
This is the chance of observing 7 or more successes in 7 trials.

The two-tail P value is 0.0156

This is the chance of observing either 7 or more successes, or 0 or fewer successes, in 7 trials.

p-value for the sign test < 0.05

Another example:



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

Control pageviews: 27,948 Control CTR: 0.1016

Experiment pageviews: 28,052 Experiment CTR: 0.1132

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

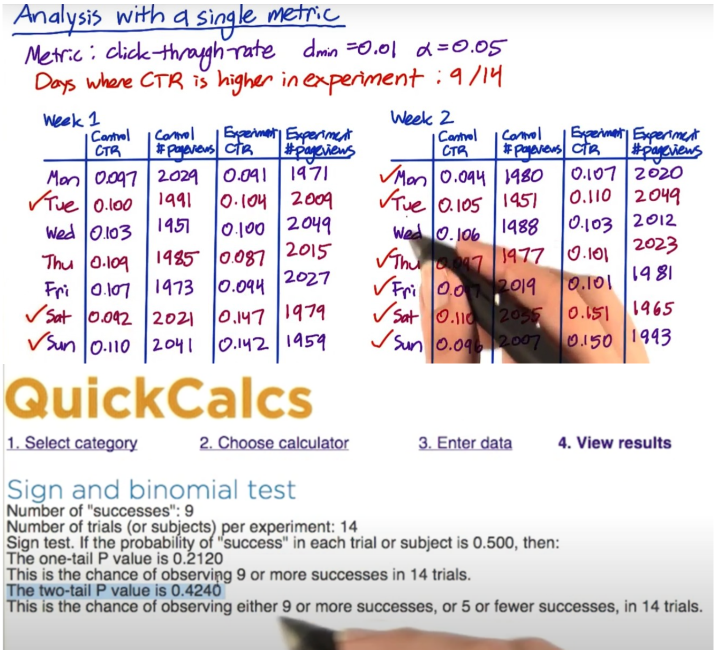
$$\frac{SE}{\sqrt{\frac{1}{27,948}+\frac{1}{28,052}}}=\frac{0.0062}{\sqrt{\frac{1}{5000}+\frac{1}{5000}}} \qquad SE=0.0026$$

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

The CI does not include 0 – indicating that the difference is at 95% level significantly different from 0. But 0.01 is included, indicating that cannot be 95% confident that the change is greater than 0.01 – i.e. the size effect we care about.

Sign test:

Two-tail p-value is 0.424, which is not significant at 0.05 level.



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
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	1959

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	2055	0.151	1965
✓Sun	0.096	2007	0.150	1993

QuickCalcs

1. Select category 2. Choose calculator 3. Enter data **4. View results**

Sign and binomial test

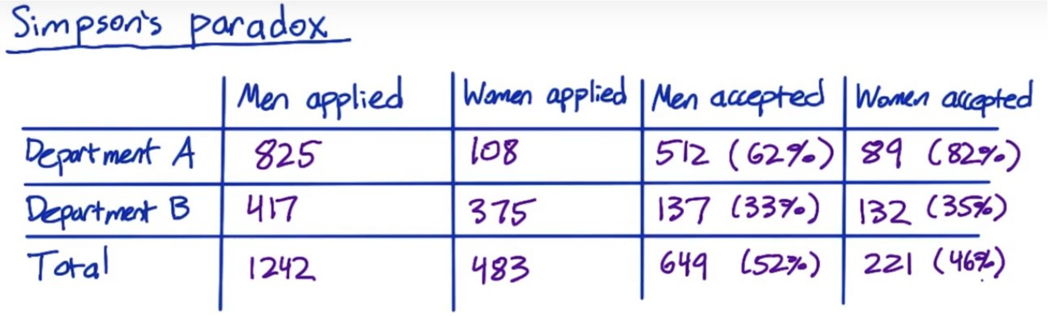
Number of "successes": 9
Number of trials (or subjects) per experiment: 14
Sign test. If the probability of "success" in each trial or subject is 0.500, then:
The one-tail P value is 0.2120
This is the chance of observing 9 or more successes in 14 trials.
The two-tail P value is 0.4240
This is the chance of observing either 9 or more successes, or 5 or fewer successes, in 14 trials.

Hypothesis on the effect size showed statistically significant results, but the sign test didn't. Why?

1. The sign test has lower power than the effect size test, which is frequently the case for nonparametric tests. That's the price you pay for not making any assumptions. So this isn't necessarily a red flag, but it's worth digging deeper and figure out what's going on.
2. In the above example, weekend click-through rates are much higher than weekdays. Size effect on weekends are a lot higher. Weekdays don't have a statistically significant difference but weekends have.

Based on this, I would recommend not launching the experiment at this point. Instead, I would dig deeper into why the change didn't affect weekday visitors. Once I understood that, I might have an idea for how to iterate on the change to help it affect more of the users.
If not, then I'd talk to the decision makers about whether a change of this magnitude on weekend traffic is worth launching.

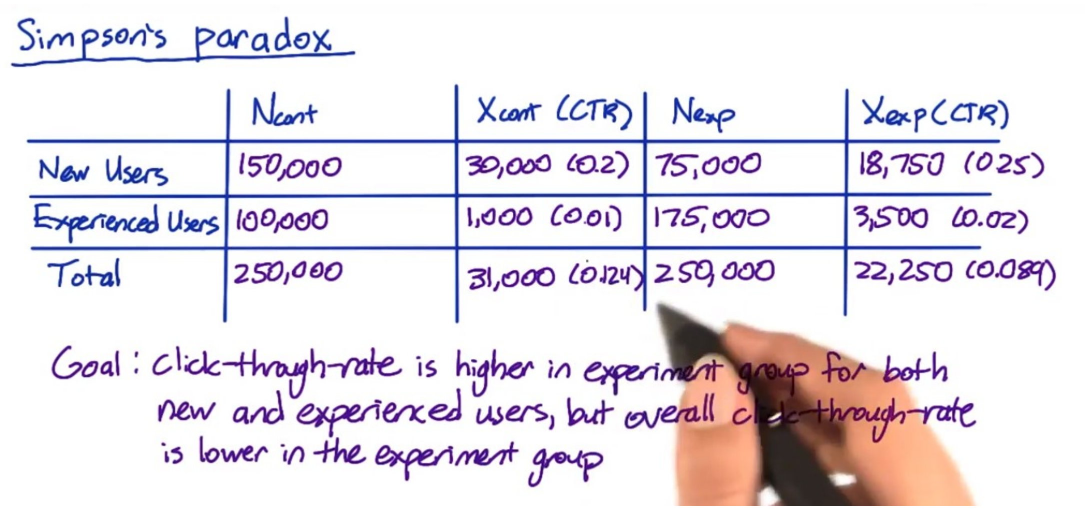
Simpson Paradox: different subgroups in the data, within each group the results are stable, but when aggregated the mix of the subgroups drive the results.



Simpson's paradox

	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%)

Women's acceptance rate is higher than men in both departments, but the overall acceptance rate is lower. This is because most of the women applied to department B, whose acceptance rate is lower than A.



Simpson's paradox

	Ncont	Xcont (CTR)	Nexp	Xexp (CTR)
New Users	150,000	30,000 (0.2)	75,000	18,750 (0.25)
Experienced Users	100,000	1,000 (0.01)	175,000	3,500 (0.02)
Total	250,000	31,000 (0.124)	250,000	22,250 (0.089)

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

CTR for experiment group is higher than control for both new users and experienced users. But for total users, control group CTR is higher. This is because control group has more new users, which has higher CTR and experienced users.

Problems:

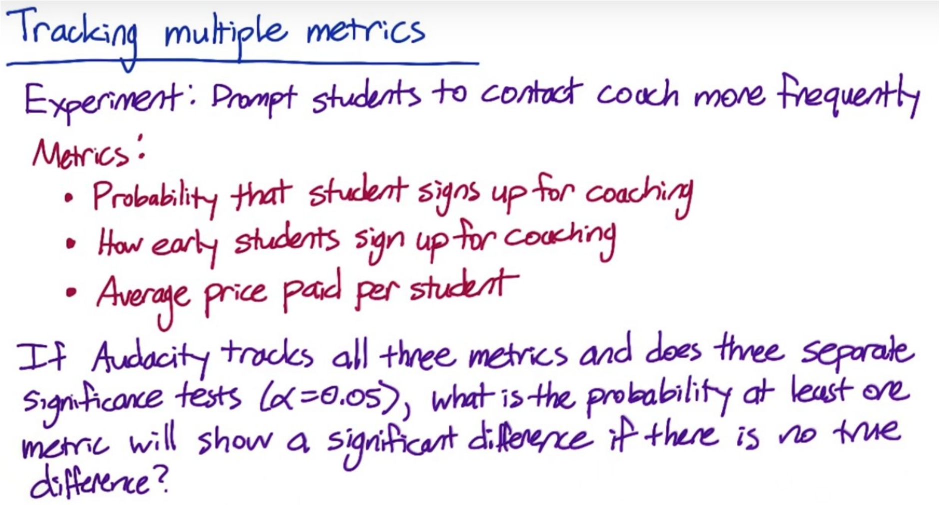
1. Why are there more page views from new users in the control group than in the experiment group? If the assignment to the control in the experiment group is random, then shouldn't the new users be evenly split between the control and experiment? And same for the experienced users.
   It should be. So this problem within the experiment setup largely affects the result. It's a good idea to make sure the number of page views is the same in the experiment group and the control group as a sanity check. Checking that breakdown across different slices could also be a good sanity check.
2. However, it's also possible to get skewed numbers like this, even if your setup is correct, if your change, or experiment, affects new users and experienced users differently.
   Suppose you're diverting based on user ID and the change makes new users generate fewer page views, for example, they refresh the page less, and experienced users generate more page views. That explains why there are more page views in the experiment group for experienced users and more page views in the control group for new users.

Therefore, although for each subgroup it seems that CTR has improved, the overall CTR was not improved and cannot say the experiment is successful. Whether it's a faulty experiment set up, or something where your change affects new and experienced users differently, you won't be able to make a valid conclusion until you understand what's going on.

Multiple Metrics

As you test more metrics, it becomes more likely that one of them will show a statistically significant result by chance. So if you're testing 20 metrics, and you have a 95% confidence level. You would expect to see one case at least that time where you got a result that says it's significant but it's only concurring by chance.

So this is a problem, but you're not sunk because it shouldn't be repeatable. That is if you did the same experiment on another day or you divide or just slices or you did some bootstrap analysis, you wouldn't see the same metric showing up as significant differences every time, it should occur randomly.



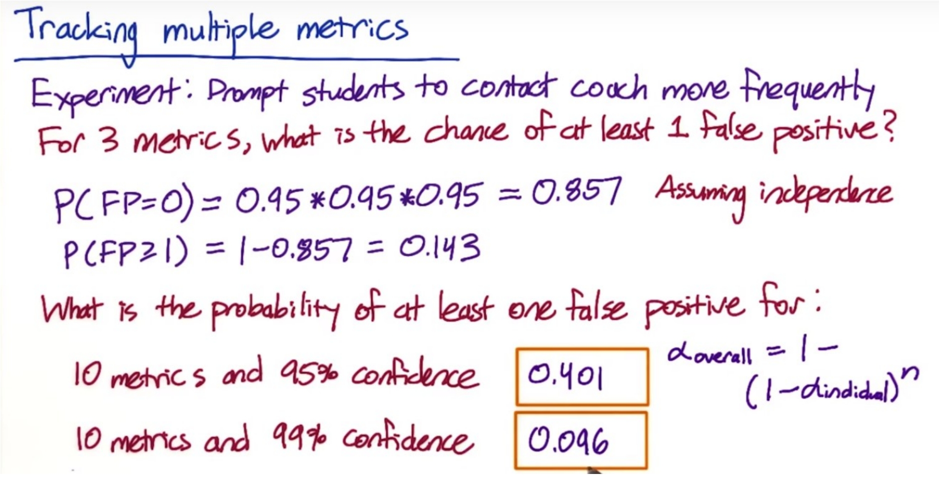
Tracking multiple metrics

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?



Tracking multiple metrics

Experiment: Prompt students to contact coach more frequently

For 3 metrics, what is the chance of at least 1 false positive?

$P(FP=0) = 0.95*0.95*0.95 = 0.857$ Assuming independence

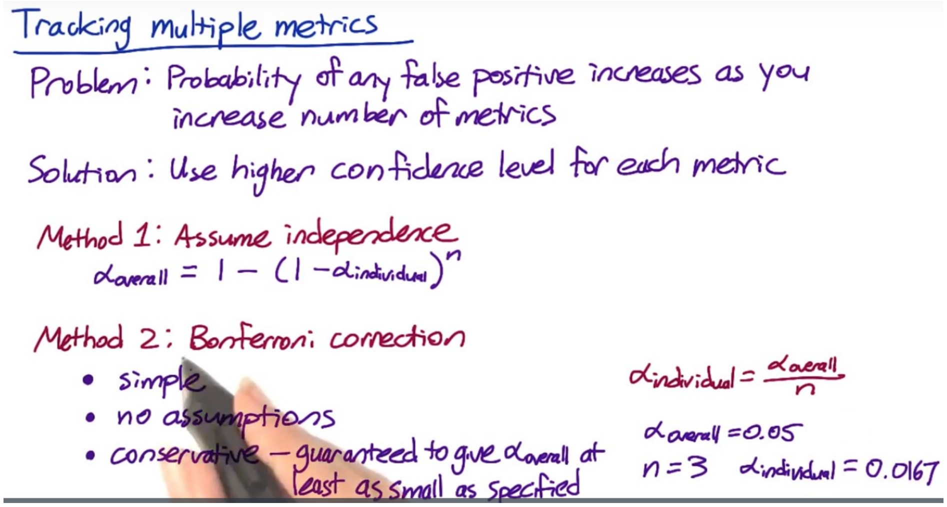
$P(FP\geq 1) = 1-0.857 = 0.143$

What is the probability of at least one false positive for:

10 metrics and 95% confidence $\boxed{0.401}$ $\alpha_{overall} = 1-(1-\alpha_{individual})^n$

10 metrics and 99% confidence $\boxed{0.096}$

I was assuming that the metrics were independent. In fact, this isn't true here. These three metrics are all related and more likely to move together. So 14.3% is an overestimate of the probability of a false positive. But assuming independence is an easy way to get a conservative estimate.



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

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

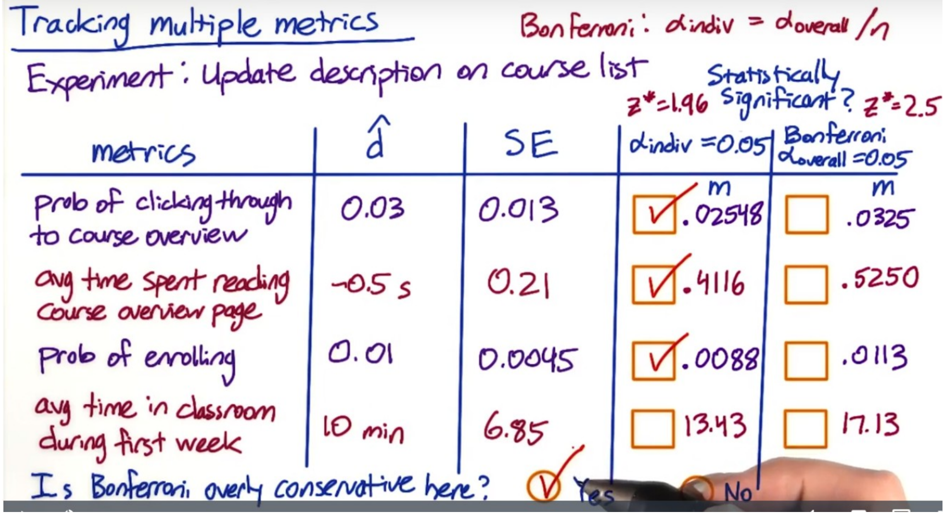
$\alpha_{individual} = \frac{\alpha_{overall}}{n}$

$\alpha_{overall}=0.05$

$n=3$ $\alpha_{individual}=0.0167$

Method 1: set up an overall alpha and use it to calculate each individual alpha.

Method 2: often will be tracking metrics that are correlated and all tend to move at the same time, in which case this method is too conservative – this results in less significant difference, and launch less experiments.



Tracking multiple metrics

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

Experiment: Update description on course list

Statistically significant?

metrics	$\hat{d}$	SE	$z^*=1.96$ $\alpha_{indiv}=0.05$	$z^*=2.5$ Bonferroni $\alpha_{overall}=0.05$
			m	m
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

In this case, the Bonferroni method is probably too conservative. If the course description was an improvement, then it makes sense that it could cause more than one of these metrics to move and they're probably more likely to move together.

Analyze the multiple metrics
Are all the related metrics moving in the same direction – e.g. click-through rate and click through probability.
Revenue per thousand queries is composed of click through rate and cost per click

Stay time on the page vs. clicks on the page – people might spend more time on clicking than staying. Need to better understand how people reacts to the changes.

Overall evaluation criteria (OEC) should be established based on an understanding of what your company is doing and what the problems are. It should balance long-term and short-term benefits. Business analysis is needed to make the decision. Once you have some candidates of OEC, you can run a few experiments to see how they steer you (whether in the right direction).

Change in metric and not others:

- Maybe you know for small changes, a change in one metric and not others might be fine.
- But for big changes, this may indicate something is wrong. Depends on your understanding of the changes itself.

Different impact across slices:

- Again need to understand the changes. Is there a bug? Have you seen this in other experiments? Is this because of different users (like or do not like the change)
- e.g. bolding works better in English/German than Chinese/Japanese. May consider using color than bolding for Chinese/Japanese.

Whether to launch an experiment or not??
1. Statistically and practically significant to justify the change?
2. Do you understand what the change can do to user experience?
3. Is it worth the investment?

Ramp up AB test

Maybe start with 1% of the traffic and divert to experiment and increase that until the feature is fully launched.

Also remove all filters to test the change on all users to understand if there is any incidental impact to unaffected users that you didn't test in your original experiment.

Gotcha: the effect might flatten out as you experiment the change – effects are not repeatable even they are statistically significant.

- Seasonality such as school season, holiday, etc.

Holdback – launch the experiment to everyone except for a small holdback who don't get the change, and you continue to compare them to the control. You will see a reverse of the impact in your experiment, and you can track that over time until you are confident that your results are repeatable. This can help track lots of seasonal or event-driven impacts.

Other things that cause the disappearing launch effect?

- Novelty effect or change aversion: as users discover or change their adoption of your change, their behavior can change and measured effect can change – can do cohort analysis.
- Pre- and post- period analysis in combination of cohort analysis to understand learning effect – i.e. how users adapt to the changes over time.

Lessons learned:
(1) Always make sure your experiment setup is correct
(2) In addition to statistical significance, you are making business decision. E.g. what if it improves for 30% and neutral for the rest? Or what if it improves for 70% but makes it worse for the 30% left? Want to launch as is or fine-tune it first
(3) overall business analysis – what's the engineering cost of maintaining the change? Are there customer support or sales issue? What's the opportunity cost? These are judgment calls which your recommendation should be based on.
(4) As noted earlier, test for all users for the incidental impact.