Notes on A/B Testing (Udacity)

Yuanzhe Li

2020-02

Contents

Notes on the A/B Testing (Udacity) course.

Lesson 1: Overview of A/B Testing

1.15 Calculating confidence interval (CTR example)

- Let's say we have $\hat{p} = \frac{X}{N} = \frac{100}{1000} = 0.1$ where $X = \#$ of users who clicked, and $N = \#$ of users.
- A *Rule of Thumb* for normality is to check $N \cdot \hat{p} > 5$ (might as well check $N \cdot (1 - \hat{p}) > 5$), otherwise use t-distribution instead of z-distribution.
- The *margen of error* $m = z_{\alpha/2} \cdot SE = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$, notice that here $SE = \sqrt{\frac{p(1-p)}{N}}$ instead of $\sqrt{np(1-p)}$ for binomial distribution, since we use the fraction or proportion of successes instead of the total number of sucesses.
- For $\alpha = 5\%$, we have $m = z_{0.025} \cdot \sqrt{\frac{0.1*0.9}{1000}} = 0.019$, and the final 95% CI is $[0.081, 0.119]$.

1.17 Null and Alternative Hypothesis, *Two-tailed vs. One-tailed tests*

The null hypothesis and alternative hypothesis proposed here correspond to a two-tailed test, which allows you to distinguish between three cases: - A statistically significant positive result - A statistically significant negative result - No statistically significant difference.

Sometimes when people run A/B tests, they will use a one-tailed test, which only allows you to distinguish between two cases: - A statistically significant positive result - No statistically significant result

Which one you should use depends on what action you will take based on the

results.

If you're going to launch the experiment for a statistically significant positive change, and otherwise not, then you don't need to distinguish between a negative result and no result, so a one-tailed test is good enough. If you want to learn the direction of the difference, then a two-tailed test is necessary.

1.19 Pooled Standard Error

- We have $X_{cont}, X_{exp}, N_{cont}, N_{exp}$, and
- Pooled sample mean $\hat{p}_{pool} = \dfrac{X_{cont} + X_{exp}}{N_{cont} + N_{exp}}$
- Pooled sample standard error $SE_{pool} = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})(\frac{1}{N_{cont}} + \frac{1}{N_{exp}})}$
- Test statistic $\hat{d} = \hat{p}_{exp} - \hat{p}_{cont}$
- Null hypothesis $H_0 : d = 0$, under which $\hat{d} \sim \mathcal{N}(0, SE_{pool})$
- For 95% confidence level ($z_{1-0.05/2} = 1.96$), if $\hat{d} > 1.96 \times SE_{pool}$ or $\hat{d} < -1.96 \times SE_{pool}$, reject the null.

1.21 - 24. Sample Size and Power

- Two types of error
  - $\alpha = P(\text{reject null} \mid \text{null True})$
  - $\beta = P(\text{not reject null} \mid \text{null False})$
  - So if sample is small, we have low α and high β, i.e., harder to identify the alternative when a difference exists. On the other hand, if sample is large, α is the same, but β is much lower, as shown below.
    - Sample size = 1000
    - Sample size = 5000
  - $1 - \beta$ is called *sensitivity* and often choose to be $> 80\%$
- Note on power
  - Statistical textbooks often define power as the sensitivity. However, conversationally power often means the probability that your test draws the correct conclusions, which depends on both α and β.
- Required sample size to achieve certain statistical power can be calculated using online calculator, in which you need to specify α, β, baseline conversion rate (null), and minimum detectable effect (alternative).
- Final notes on how typer I & II confidence level and detectable difference d_min can determine the required sample size together is as follows
- Examples of factors that affect the required sample size are as follows:

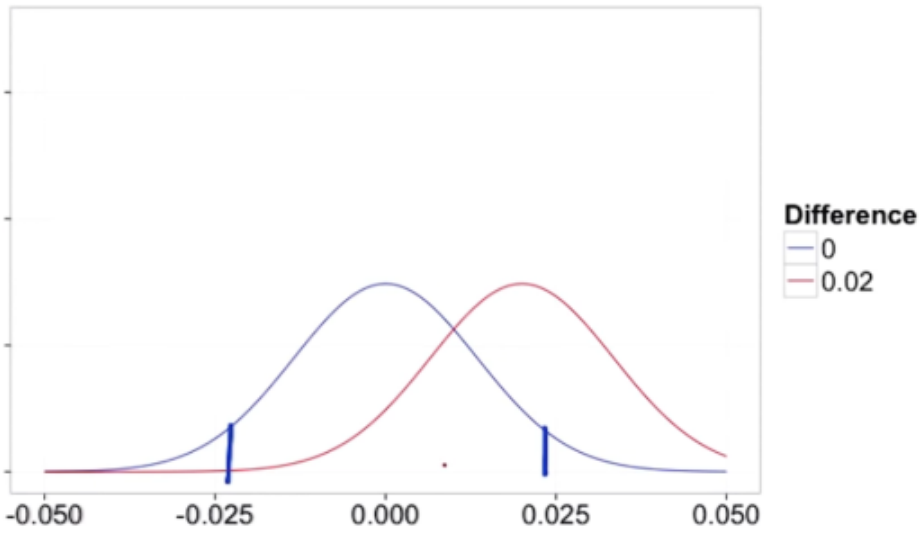


Figure 1: small_sample

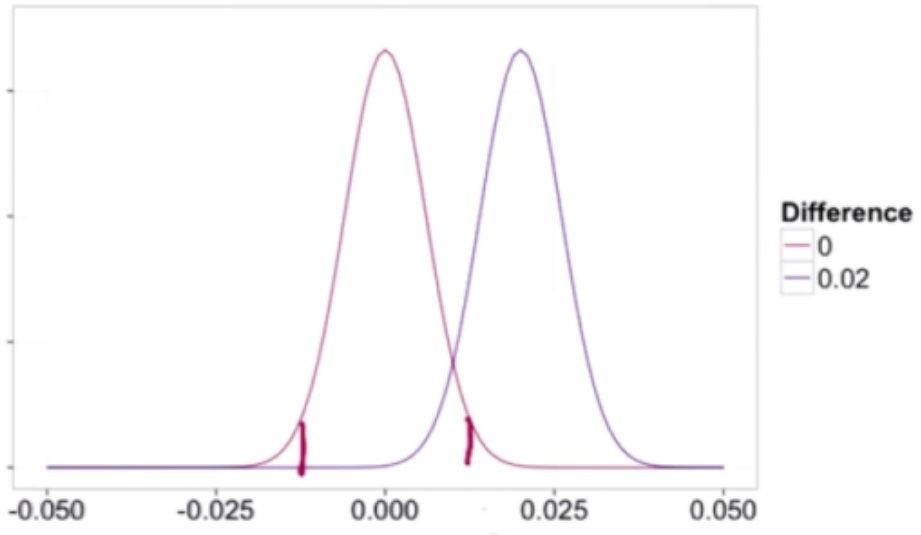


Figure 2: large_sample

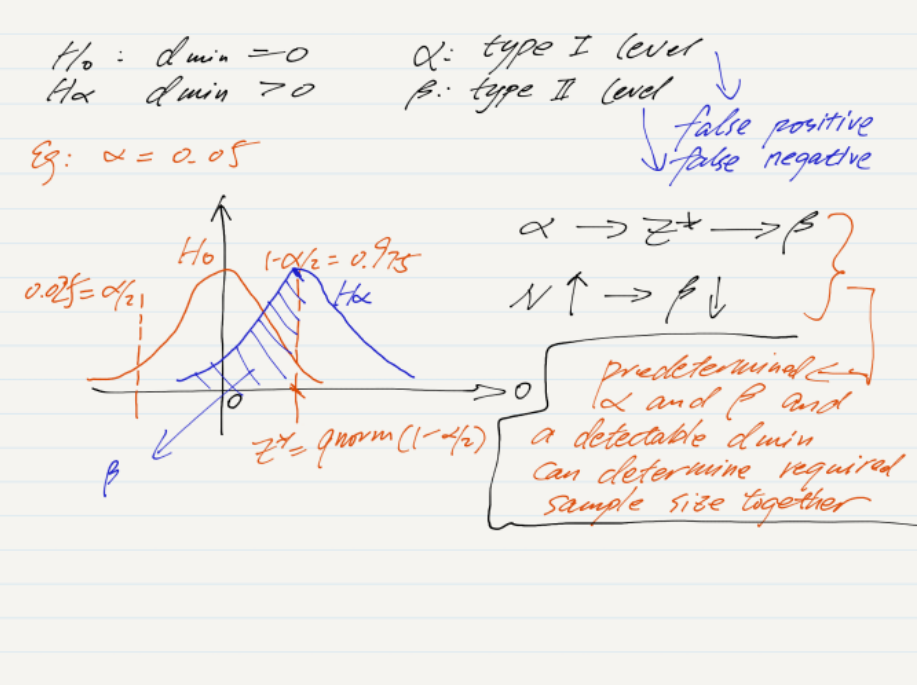


Figure 3: notes on required sample size

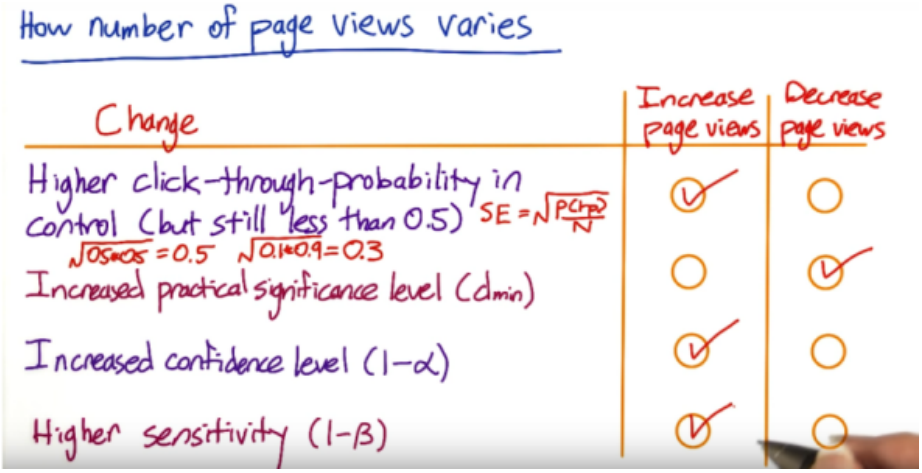


Figure 4: how many page views varies

1.25 Pooled Example

An pooled example is shown below, notice how the d_{min} works (need the lower bond of the $1-\alpha$ level CI $> d_{min} = 0.02$)

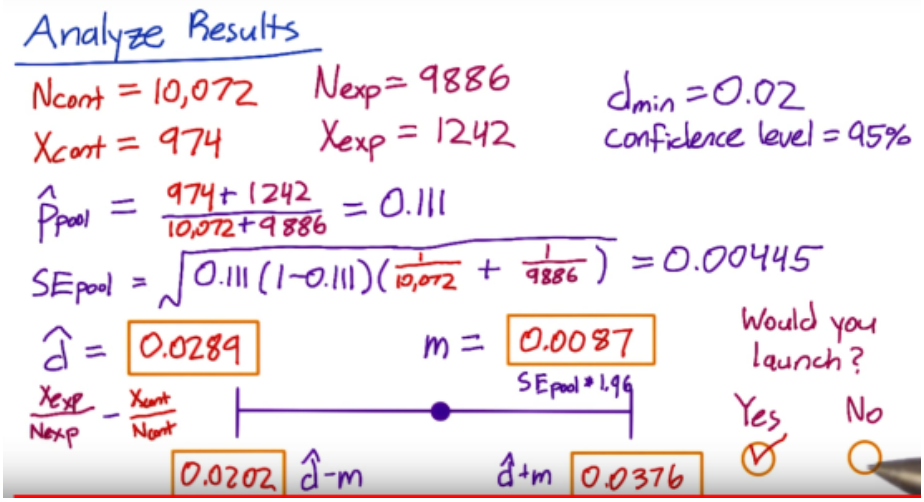


Figure 5: pooled example

1.26 Confidence Interval Case Breakdown

Shown below is the how we should consider the decision under varying CI and d_{min} cases

Lesson 2: Policy and Ethics for Experiments

2.1 - 2.7. Four Principles

IRB's four main principles to consider when conducting experimentats are: - **Risk**: *what risk is the participant undertaking?*. The main threshold is whether the risk exceeds that of *"minimal risk"*. Minimal risk is defined as the probability and magnitude of harm that a participant would encounter in normal daily life. - **Benefit**: *what benefits might result from the study?* - **Choice/Alternatives**: *what other choices do participants have?* - **Privacy/Data Sensitivity**: *what data is being collected, and what is the expectation of privacy and confidentiality?* - How sensitive is the data? - What is the re-identification risk of individuals from the data?

2.8 Accessing Data Sensitivity

An example of data sensitivity assessment is shown below

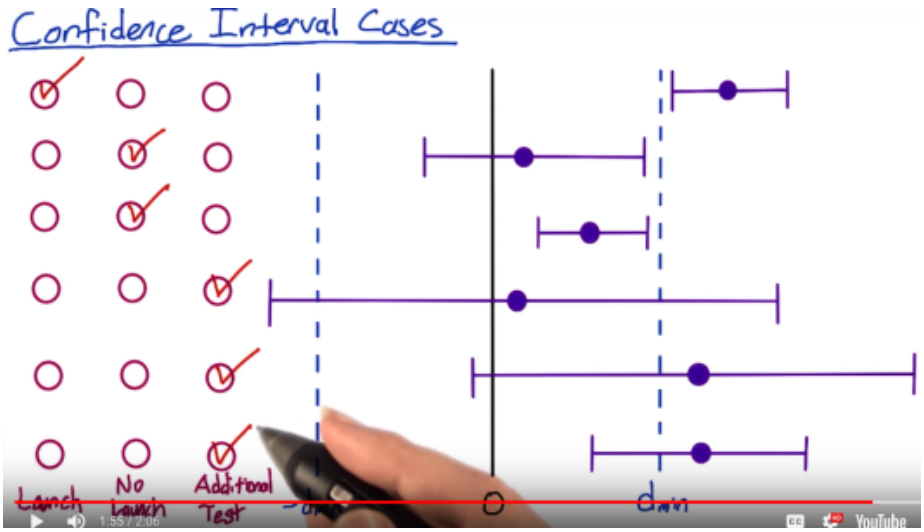


Figure 6: CI breakdown

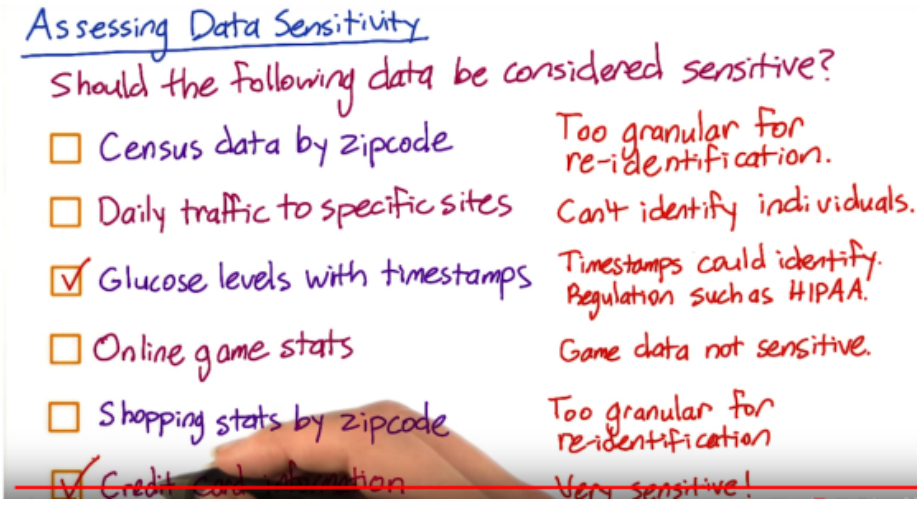


Figure 7: accessing data sensitivity

2.10 Summary of Principles

- It's a grey area whether internet studies should be subject to IRB review or not and whether informed consent is required.
- Most studies face the bigger question about data collection with regards to identifiability, privacy, and confidentiality / security.
  - Are participants facing more than minimal risk?
  - Do participants understand what data is being gathered?
  - Is that data identifiable?
  - How is the data handled?

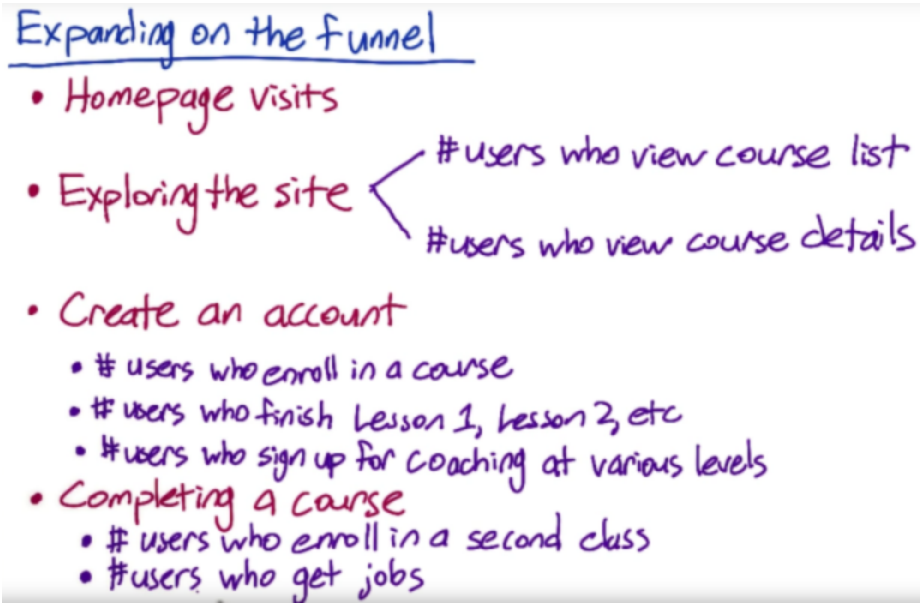
Lesson 3: Choosing and Characterizing Metrics

3.2 - 3.3 Metric Definition Overview

- *Invariant Checking*: metrics shouldn't change across experiment and control
- *Evaluation*: what do we want to use the metrics for?
  - At the evaluation stage, it's better to settle on one single objective that multiple departments within the company would most likely agree on.
  - If mutlple metrics are available or equally important, we can create a composite metric, e.g., something called objective function or OEC (Overall Evaluation Criterion, a term created by Microsoft).
  - Composite metric is less preferred, as it is better to come up with a less optimal metric that works for a suite of A/B tests than to come up with a perfect metric but only for a single test.

3.5 Refining the Customer Funnel

An example of defining metrics for Udacity - Refining the customer funnel



- High-level metrics

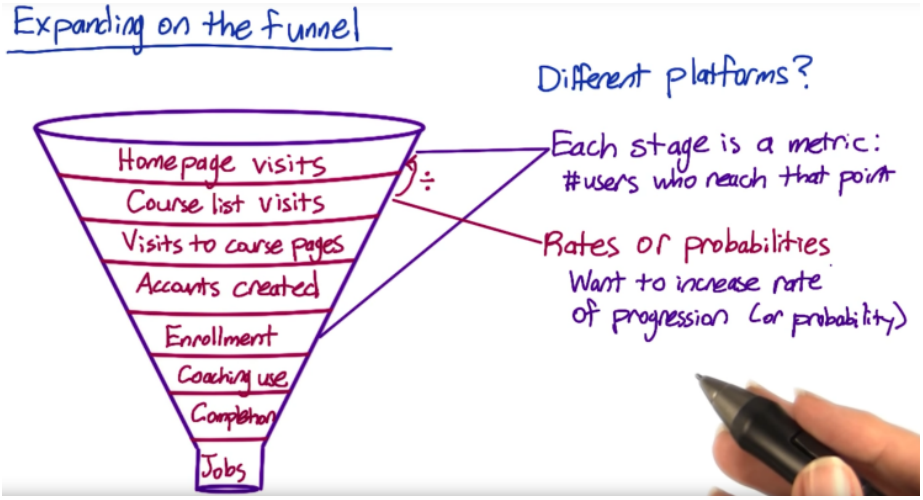


Figure 8: Metrics for funnels

3.6 - 3.7 Quizes on Choosing Metrics

- How to choose metrics for different tests
- Difficult metrics
  - Don't have access to data, e.g.,

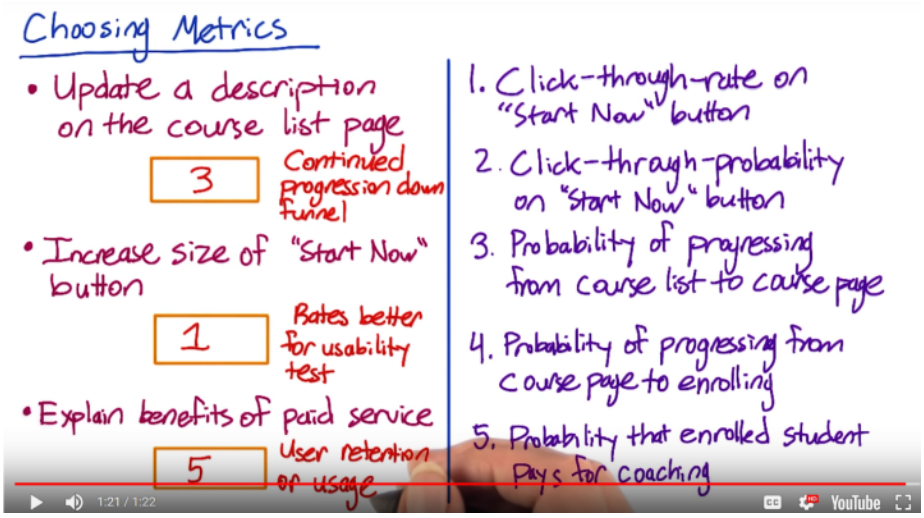


Figure 9: choosing metrics

    - Amazon wants to measure average happiness of shoppers
    - Google wants to measure probability of user finding information via search
  - Takes too long to measure, e.g.,
    - Udacity measures the rate of customers who completed the 1st course returning for 2nd one.

3.8 Other techniques for defining metrics

- External data
- User experience research, surveys, focus groups
- Retrospective analysis helps detect correlations for us to develop theories.
- For details of additional techniques for defining metrics, see `materials/define_metrics_additional_techniques.pdf`.

3.10 - 11 Techniques to Gather Additional Data and Examples

- Techniques for gather additional data
- Udacity example
- Examples where data is hard to get

3.13 Metric Definition: Click Through Example

- Metric definition

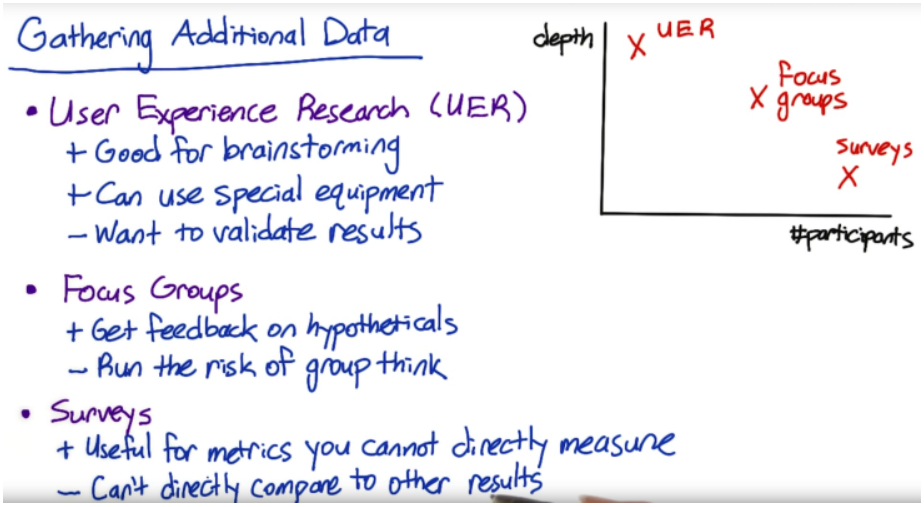


Figure 10: 3.10 techniques for getting additional data

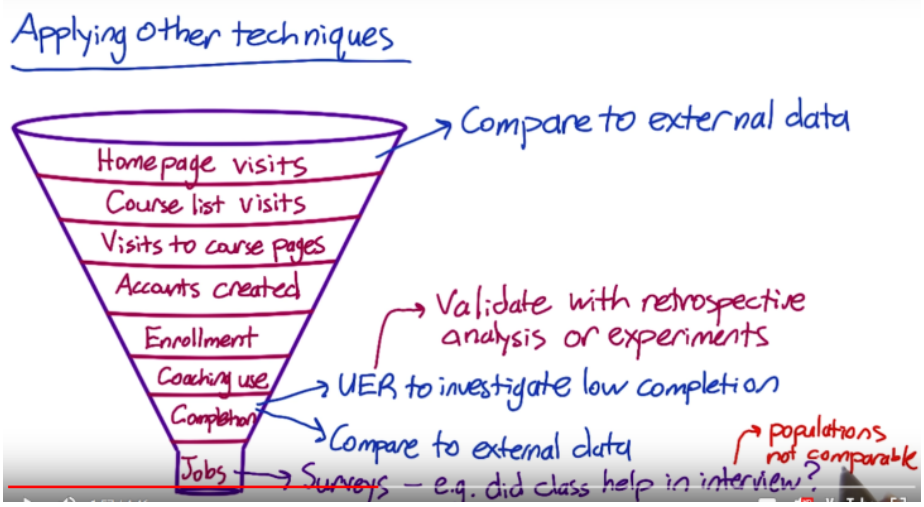


Figure 11: 3.11 gather data - udacity

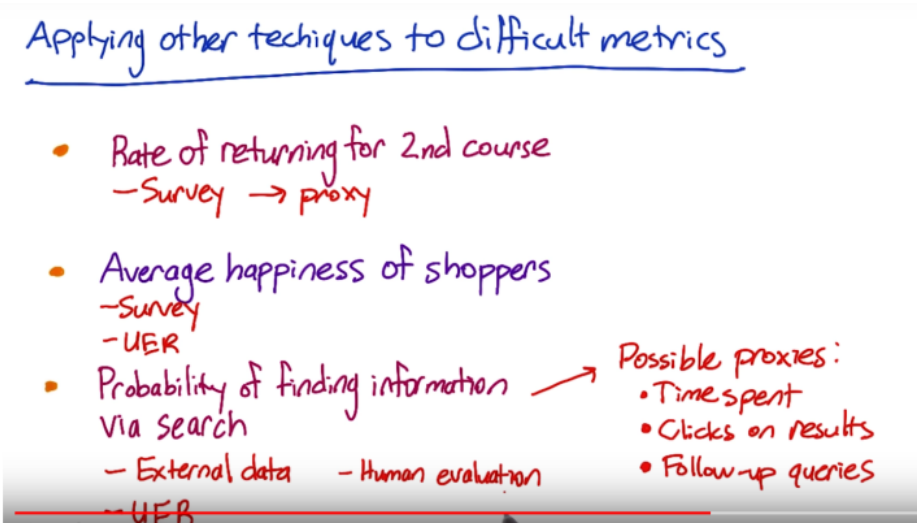
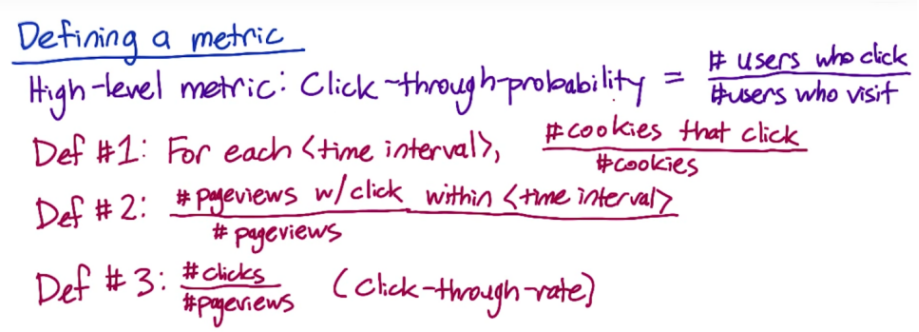
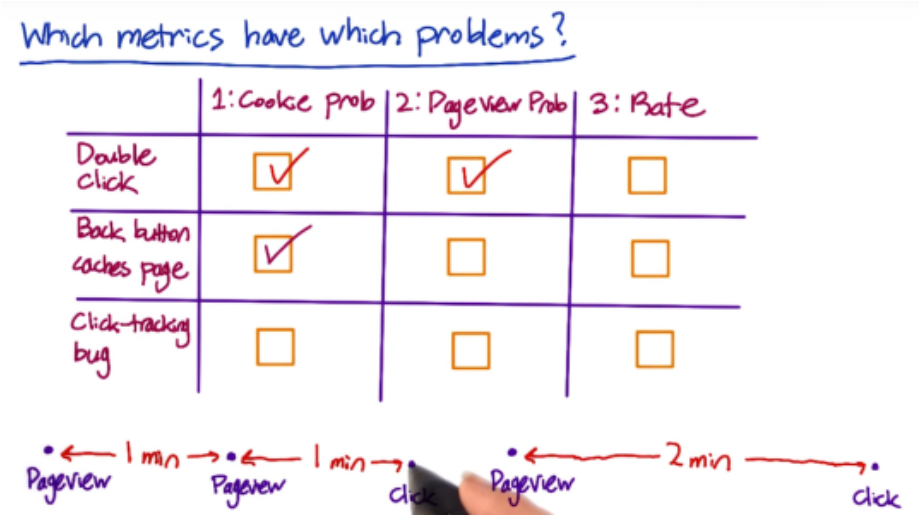


Figure 12: 3.12 when there is no data





3.16 - 3.17 Summary Metrics

- Categories of summary metrics
  - Sums and counts.
    - e.g., # users who visited page
  - Means, medians, and percentiles
    - e.g., mean age of users who completed a course or
    - median latency of page load
  - Probabilities and rates
    - Probability has 0 or 1 outcome in each case
    - Rate has 0 or more
  - Ratios
    - e.g., $\frac{P(\text{revenue-generating click})}{P(\text{any click})}$

3.18 - 3.19 Sensitivity and Robustness

- We want summary metrics to be sensitive on things we care and robust on things we don't care.
- Example: choose summary metric for latency of a video
  - Use *retrospective analysis* to check robustness. For example, if we plot distribution for similar videos and find the 95th and 99th percentiles of load time has noticeable variations between videos, those two metrics may not be **robust enough**.
  - We can also look at *experimental data*. For example, if we plot distribution of load time for videos with increasing resolution, and find that the median and 80th percentile is not affected by resolution, only the 85/90/95-th percentiles are increasing. This means that median and 80th percentile may not be sensitive enough.

3.20 Absolute Versus Relative Differences

- Usually start with absolute differences when we don't know the metric well.
- Using relative difference means we might be able to stick with the same significance boundary and not need to worry about seasonality factors (e.g., think about CTR for shopping websites)

3.21 - 3.22 Variability

- To calculate a confidence interval, we need
  - Variance (or standard deviation)
  - Distribution
- For Binomial distribution
  - $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$
  - Margin of error $m = z^* \cdot SE$, where z-score is derived from the standard normal distribution, as binomial approaches to normal when N is large.
- Distribution and estimated variance of some common metrics are as follows

Type of metric	distribution	estimated variance
probability	binomial (normal)	$\frac{\hat{p}(1-\hat{p})}{N}$
mean	normal	$\frac{\hat{\sigma}^2}{N}$
median/percentile	depends	depends
count/difference	normal (maybe)	Var(X) + Var(Y)
rates	poisson	$\bar{X}$
ratios (e.g., $\frac{\hat{p}_{exp}}{\hat{p}_{control}}$)	depends	depends

- Some summary metrics may be harder to analyze. E.g., median could be non-normal if data is non-normal (e.g., latency with bimodal distribution shown below)
- Example: calculate the 95% CI for a mean with N = [87029, 113407, 84843, 104994, 99327, 92052, 60684]
  - $\bar{N} = 92,052, \hat{sigma} = 17,015$
  - $SE = \frac{\hat{\sigma}}{\sqrt{7}} = 6,430$
  - *Margin of error* $m = z^* \cdot SE = 1.96 \cdot SE = 12,605$
  - The CI is $79,158$ to $104,367$

3.24-25 Empirical Variability

- Uses of A/A tests

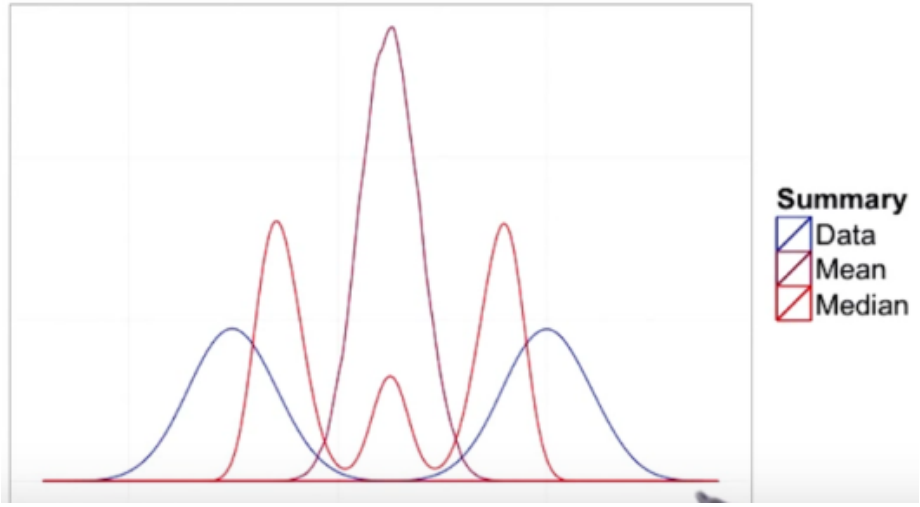


Figure 13: bimodal distribution of latency

  - Compare results to what you expect (sanity check)
  - Estimate variance and calculate confidence
  - Directly estimate confidence interval

- A CTR example
  - Spreadsheet
  - Estimate variance and calculate CI using pooled results

Calculating variability empirically

Estimate variance and calculate confidence interval:

Since we expect a normal distribution:

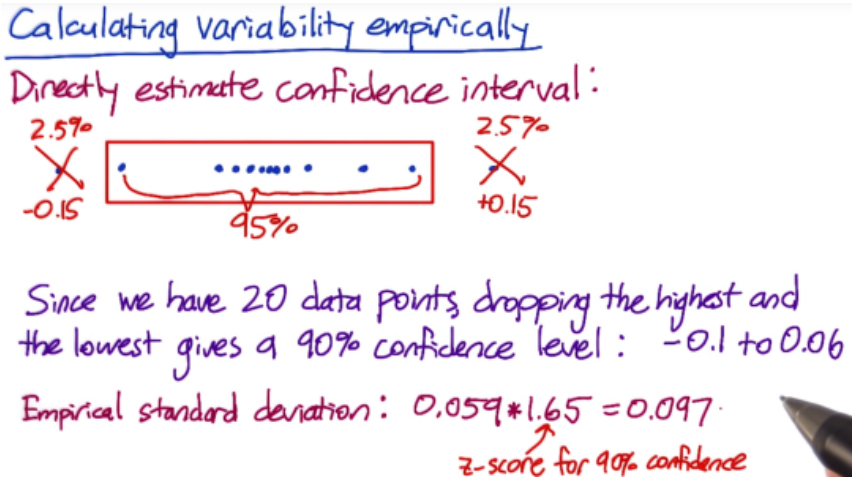
$m = SD \cdot z^*$

$= 0.059 \cdot 1.96 = 0.116$ empirically

Analytically: $SE = \sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})\left(\frac{1}{N_{cont}} + \frac{1}{N_{exp}}\right)}$

Slightly different margin of error for each experiment

  - Directly estimate confidence interval from empirical distribution



- We can also use *bootstrap* to generate multiple samples/metrics to estiamte the variability.

Lesson 4: Designing an Experiment

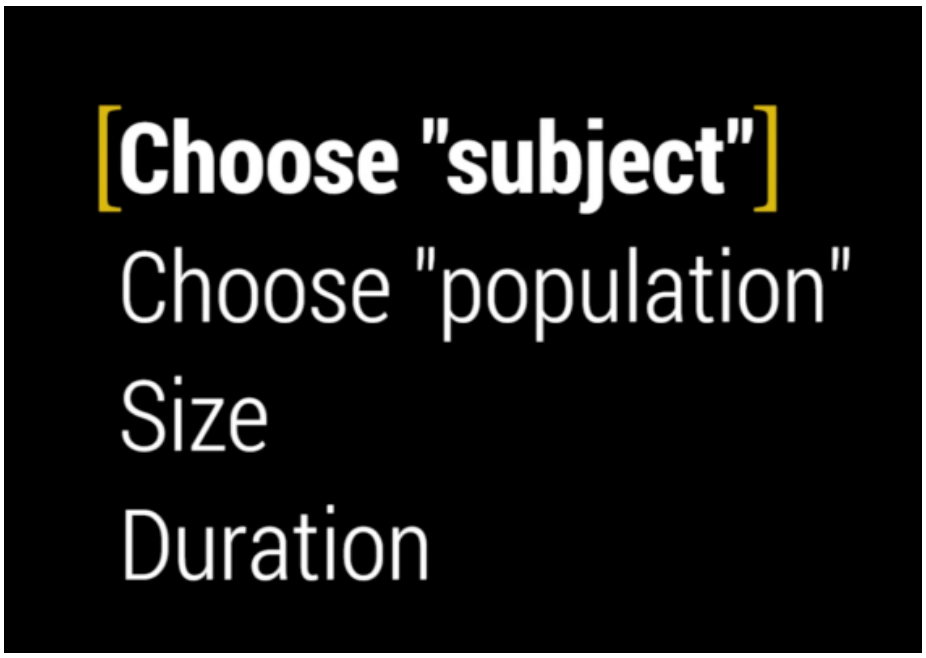


Figure 14: Outline of lesson 4

4.2 - 4.3 Unit of Diversion Overview

- *Unit of diversion* is how we define what an individual subject is in the experiment.
- Commonly used:
  - User id
    - Stable, unchanging
    - Personally identifiable
  - Anonymous id (cookie)
    - Changes when you switch browser or device
    - Users can clear cookies
  - Event
    - No consistent experience
    - use only for non-user-visible changes
- Less common:
  - Device id
    - only available for mobile
    - tied to specific device
    - unchangeable by user
  - IP address
    - changes when location changes
- Example

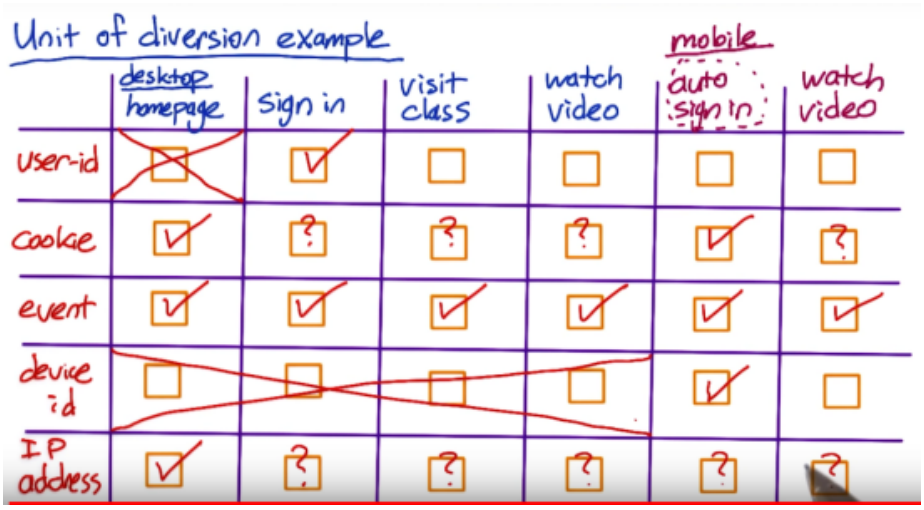


Figure 15: Example of unit of diversion

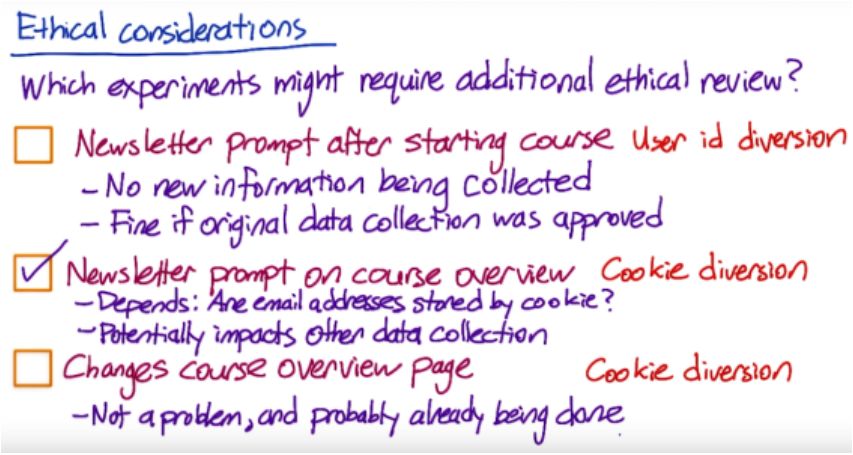
4.4 - 4.5 Consistency of Diversion

- First principle of choosing unit of diversion is to make sure users have consistent experience.
- If the customer wouldn't be likely to notice the change, we might want to start with event-based experiement. If learning effect is detected later, we can switch to a cookie-based experiment.
- Example

Figure 16: Example of consistency needed from unit of diversion

4.6 - 4.7 Ethical Considerations

- An exmaple is as follows.
  - Notice that only the second case requires additional sthical review/consent from the user because it might comprimise the anonimity of cookie-based data.



4.8 - 4.9 Unity of Analysis vs. Diversion

- Unit of analysis is basically whatever your denominator of the analysis is.
- When unit of analysis and unit of diversion is not the same, the empirical variability of the metric can be significantly larger than the analytical one. See Overlapping Experiment Infrastructure:More, Better, Faster Experimentation for more details about the following example.

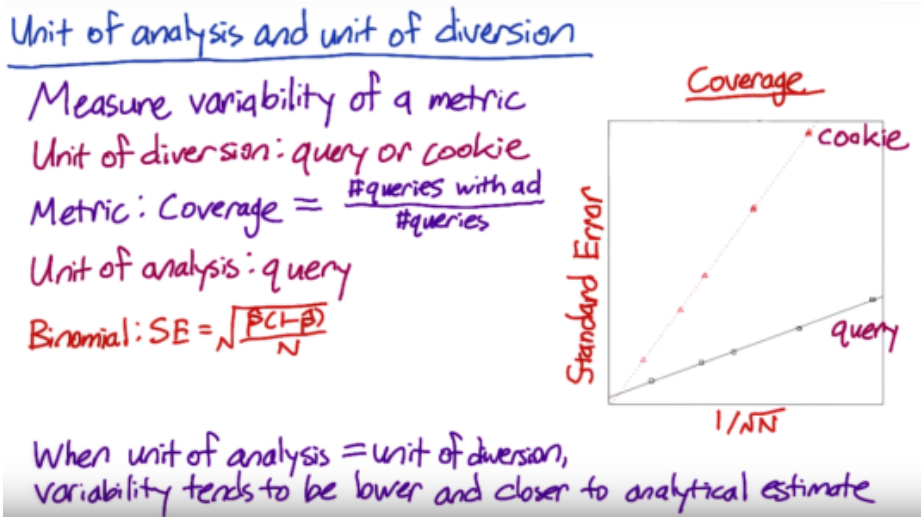


Figure 17: Unit of analysis versus diversion example

- The quiz example.

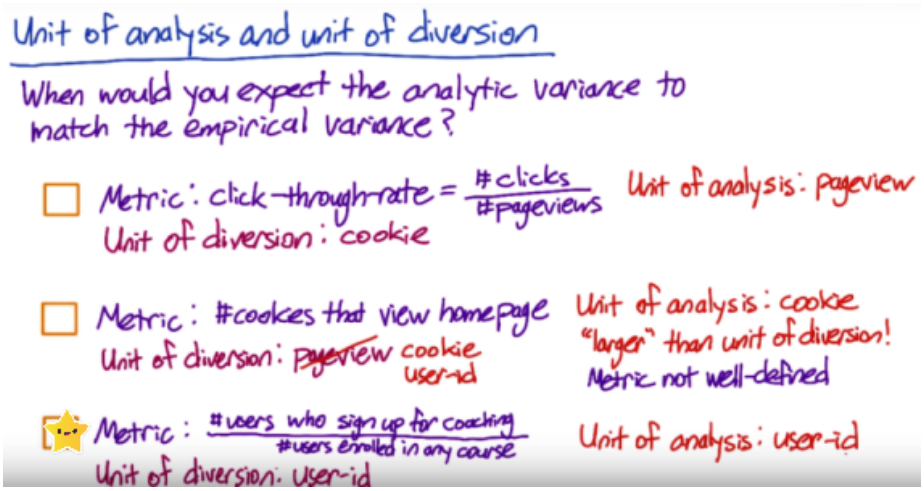


Figure 18: Unit of analysis versus diversion - quiz

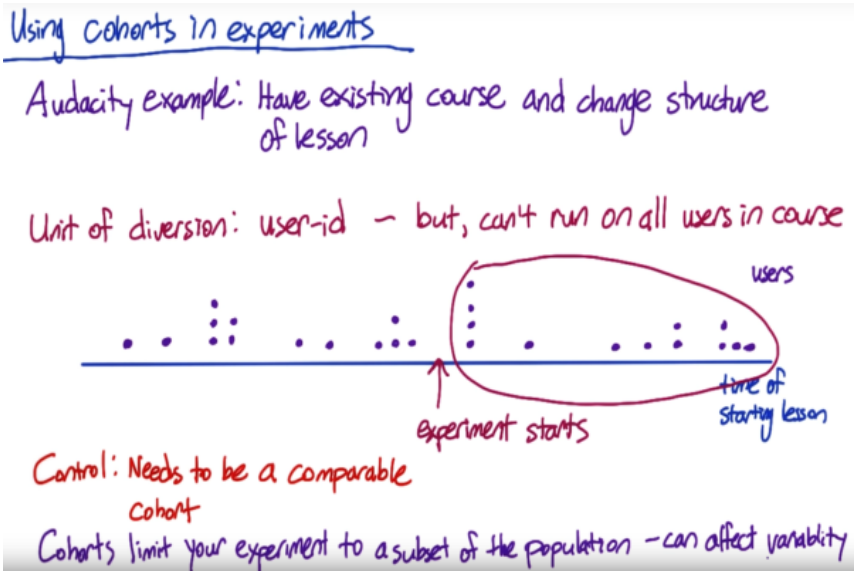
4.10 Inter- vs. Intra-User Experiments

- Interleaved experiments

In an interleaved ranking experiment, suppose you have two ranking algorithms, X and Y. Algorithm X would show results $X_1, X_2, \ldots X_N$ in that order, and algorithm Y would show $Y_1, Y_2, \ldots Y_N$. An interleaved experiment would show some interleaving of those results, for example, $X1, Y_1, X_2, Y_2, \ldots$ with duplicate results removed. One way to measure this would be by comparing the click-through-rate or -probability of the results from the two algorithms. For more detail, see Large-Scale Validation and Analysis of Interleaved Search Evaluation.

4.11 - 4.13 Target Population, Cohort

- Using cohorts in experiments
  - When to use a cohort instead of a population:
    - Looking for learning effects
    - Examining user retention
    - Want to increase user activity
    - Anything requiring user to be established
  - Audacity example:



4.16 - 4.18 Sizing Examples

- How variability affects sizing

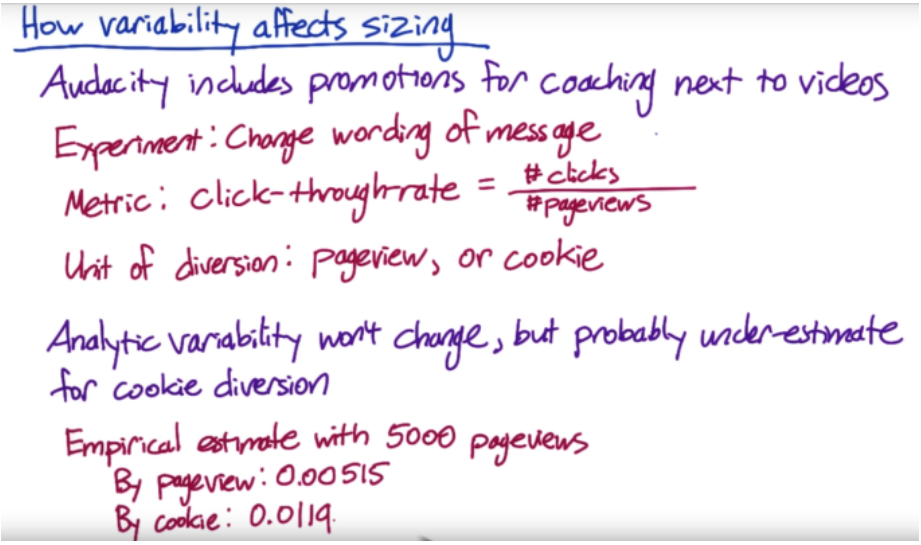


Figure 19: szing example 1

- See the codes/empirical-sizing.r for code example.
- Quiz

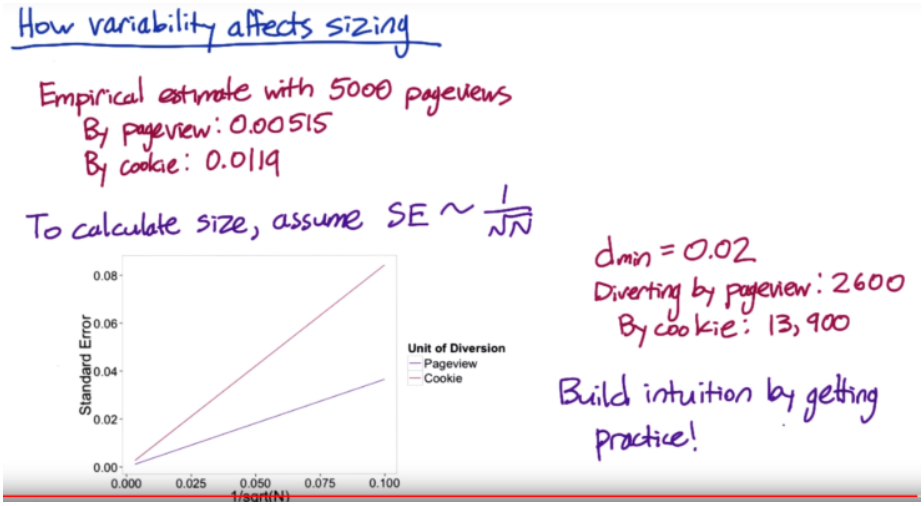


Figure 20: szing example 2

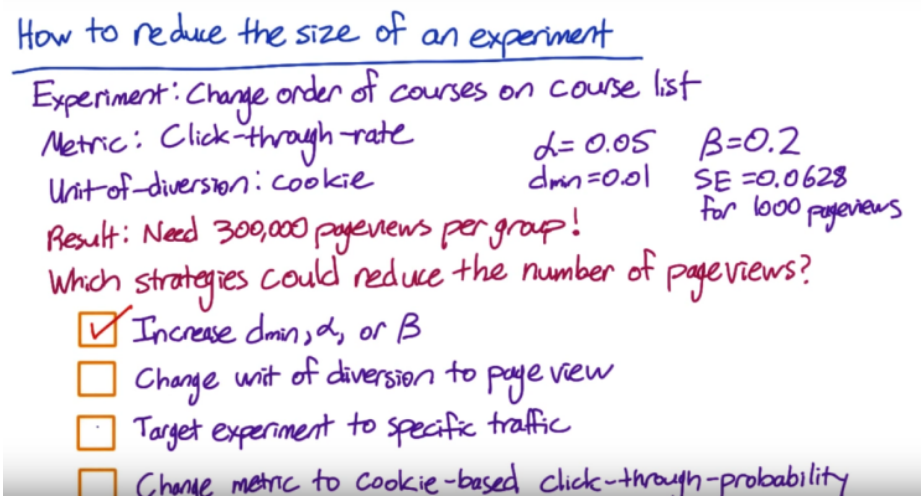


Figure 21: sizing quiz 1

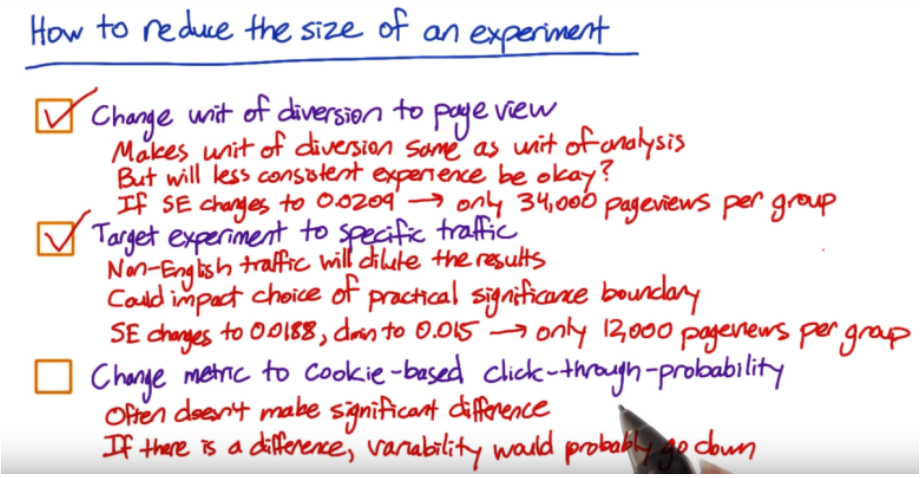


Figure 22: sizing quiz 2

4.20 - 22. Duration vs. Exposure

- Example
- When to limit exposure - quiz
  - The rule of thumb is to think about what if the worst possible impact if everything goes wrong.

4.23 Learning Effects

- Change aversion vs. novalty effect
- To measure learning effect, we need a stateful unit of diversion like a cookie or a user ID
- Better use a cohort as opposed to just a population to measure the effect of dosage (e.g., how frequent a subject sees the change)
- Risk vs. duration
- Use A/A test is useful in both pre- and post- experiment.

Lesson 5: Analyzing Results

Outline of this section - Sanity Checks - Single Metric - Multiple Metrics - Gotchas

5.1 - 5.7 Sanity Checks (invariant metrics)

- Check invariant metrics

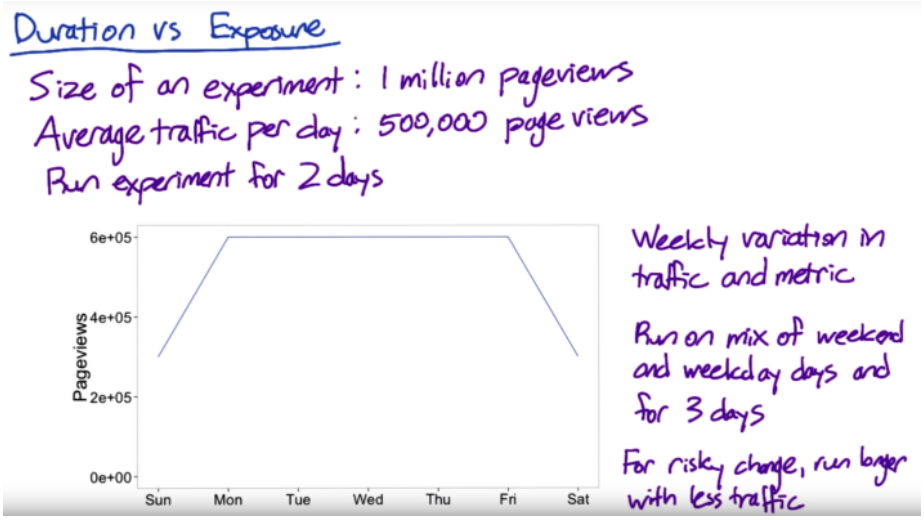


Figure 23: Duration vs exposure: example

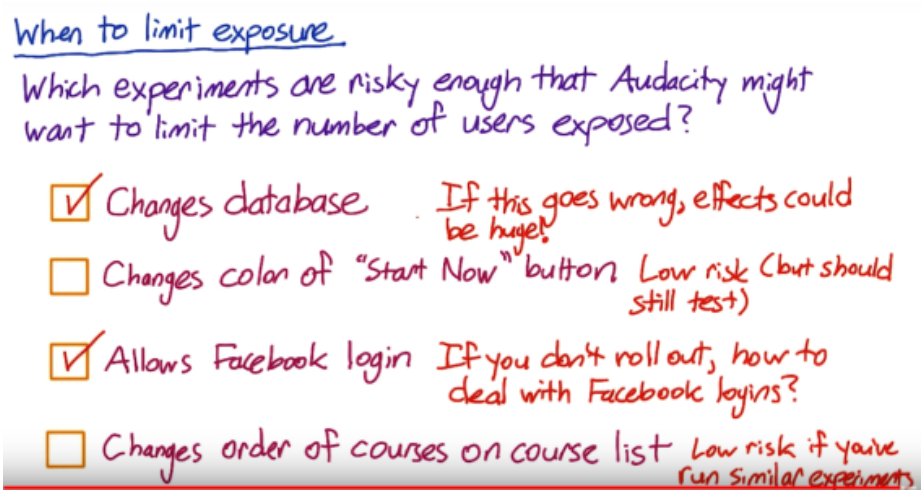


Figure 24: When to limit exposure

- Check *population sizing metrics* to make sure control and experiment groups are comparable
- Check actual invariant metrics

- Quizes
  - population sizing metrics

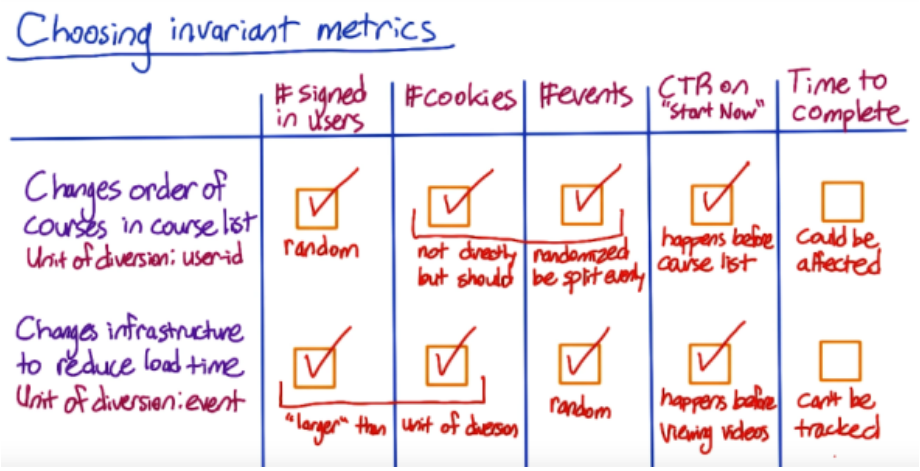


Figure 25: Quiz - 5.3 choosing population sizing metrics

- invariant metrics

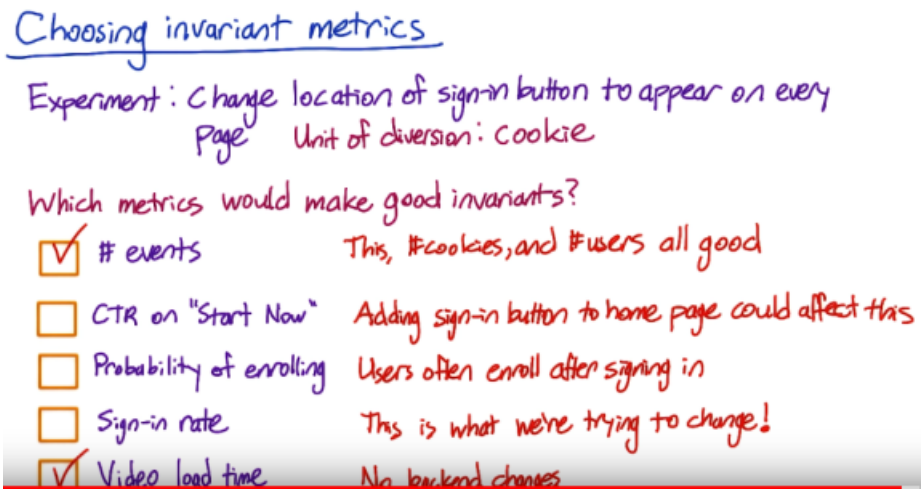


Figure 26: Quiz - 5.4 invariant metrics

- Checking invariants

Checking invariants

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

Run experiment for 2 weeks.
Unit of diversion: cookie

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

Write the steps you would take in plain English.
Then discuss on the forums and check this box: ☑

Checking invariants

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

Run experiment for 2 weeks.
Unit of diversion: cookie

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

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

Run experiment for 2 weeks.
Unit of diversion: cookie

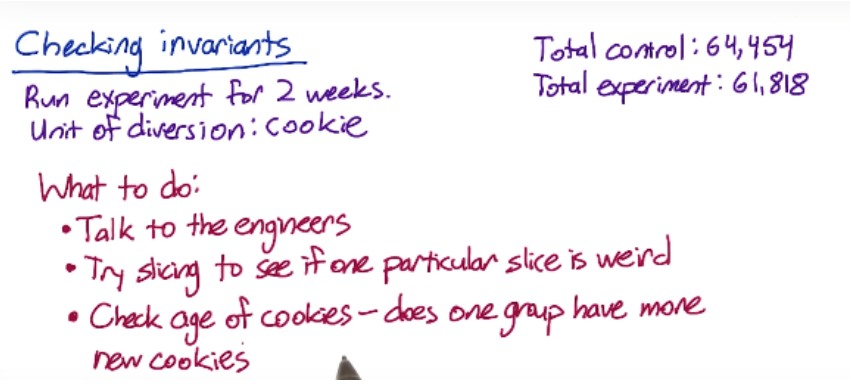
More cookies in control:

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	4816	4741	0.504
Sat	3411	2939	0.537
Sun	3478	3025	0.532

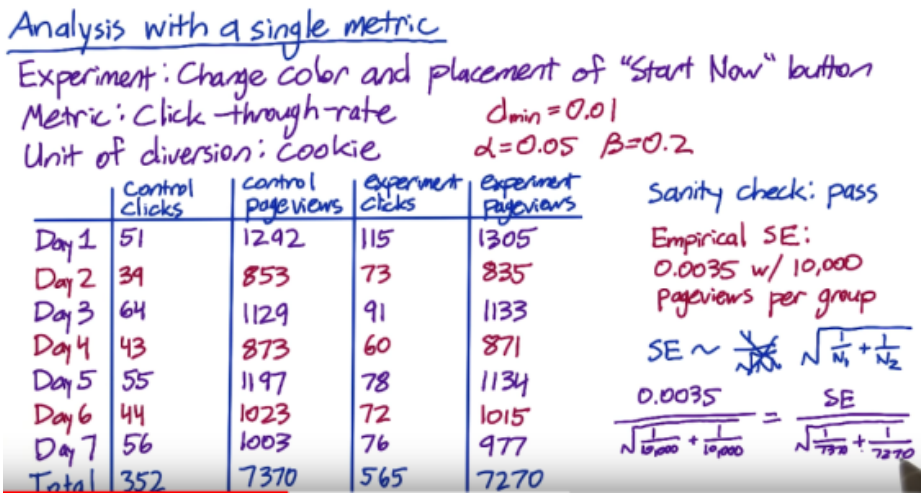


5.8 - 5.9 Single Metric

- What not to do if your results aren't significant

Carrie gave some ideas of what you can do if your results aren't significant, but you were expecting they would be. One tempting idea is to run the experiment for a few more days and see if the extra data helps get you a significant result. However, this can lead to a much higher false positive rate than you expecting! See the post (How Not To Run an A/B Test) for more details. Instead of running for longer when you don't like the results, you should be sizing your experiment in advance to ensure that you will have enough power the first time you look at your results.

- Example



	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

Figure 27: Quiz 5.9

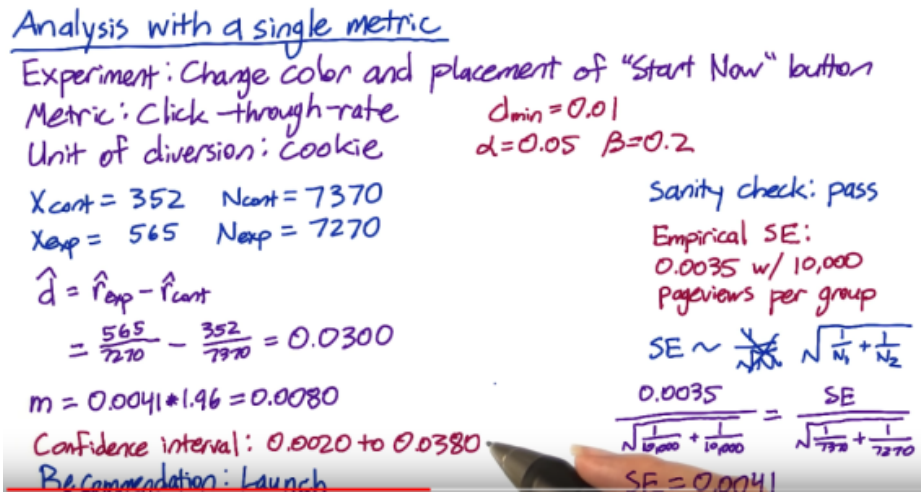


Figure 28: Quiz 5.9 - 2

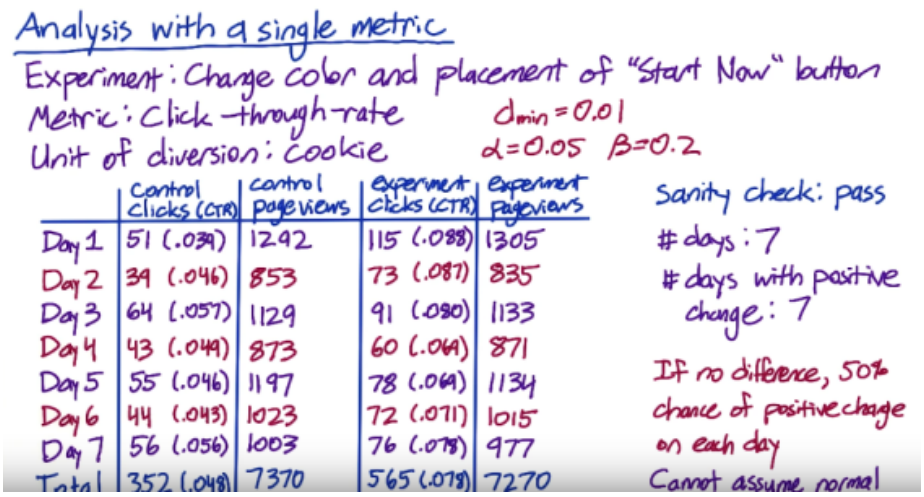
Notice that the confidence interval does not include the detectable difference d_{min}.

  - Test the probability of observaing 7 successes out of 7 experiments with success rate of 0.5. The two-tail P-value is $0.0156(< \alpha = 0.05)$, which is the probability of observing < 0 or > 7 successes in 7 experiments. Based on this, it's highly unlikely that the positive changes of CTR in the experiment group is due to chance, so we recommend launch.

- Another example
  - Notice that the CI includes d_{min} so we cannot recommend to launch
  - 9 successes out of 14 days, with a two-tail P value of 0.4240 we cannot reject the null that this is due to pure chance
  - Overall the effect size (in lifting CTR) shows significant result, but the sign test failed. Digging deeper into the day-by-day data we further observed that the effect is more significant for weekends and not significant for weekdays.

5.10 - 11. Simpson's Paradox

- An example of Simpson's paradox in A/B test is when your results within new/experienced user groups are consistent, but the aggregated result in the total population shows the reverse.



	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

Figure 29: Quiz 5.9 - 3

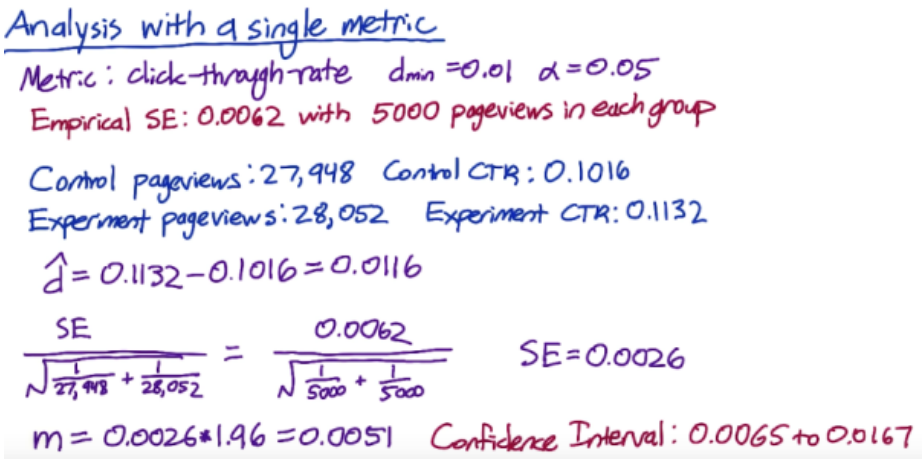


Figure 30: Quiz 5.9 - 4

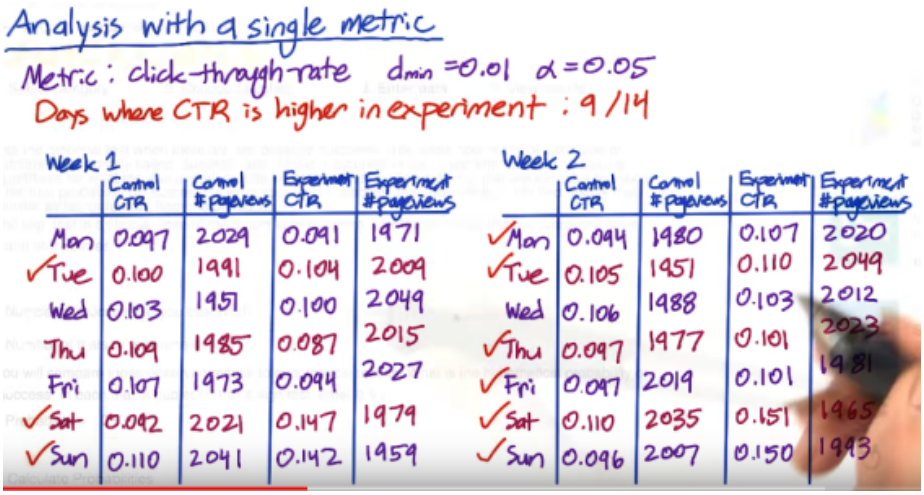


Figure 31: Quiz 5.9 - 5

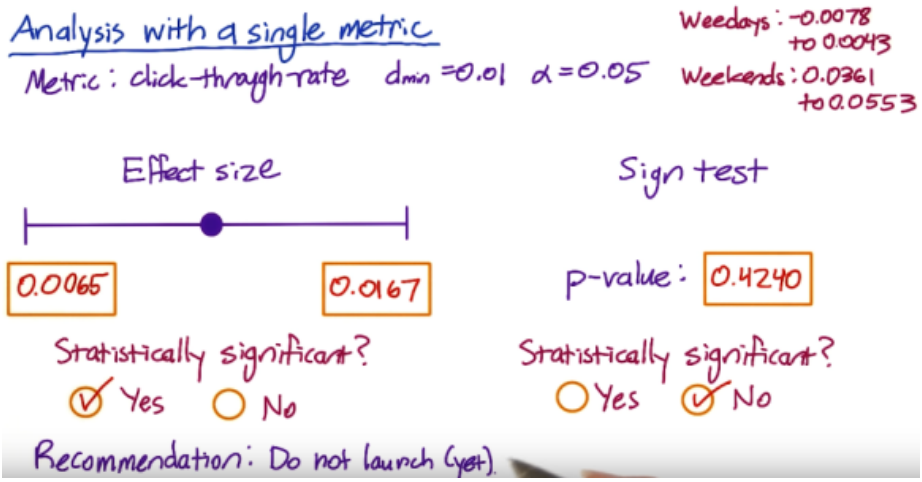


Figure 32: Quiz 5.9 - 6

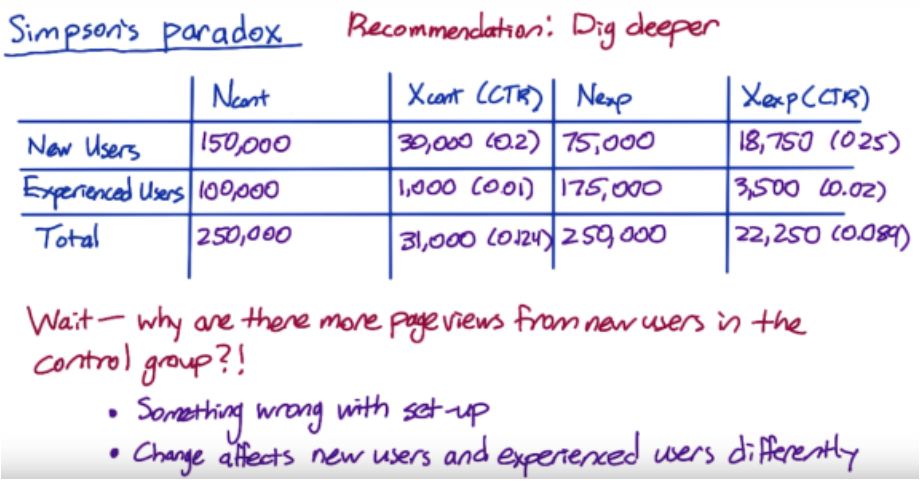


Figure 33: 5-11: simpson's paradox

5.12 - 5.15. Multiple Metrics

- Bonferroni correction can guarantee the overall FP rate $\alpha_{overall}$ by controlling individual FP rate $\alpha_{individual} = \alpha_{overall}/m$. However, it is too conservative.
- A audacity example where Bonferroni is too conservative. In the example below (Z^* is the critical value corresponding to the confidence level, and m is the margin of error), three metrics that showed significant difference individually would be rejected if we use Bonferroni to keep $\alpha_{overall}$ at the same level (i.e., 0.05), which is probably too conservative.
- Practical recommendations to counter the conservatism of Bonferroni correction includes:
  - Rigorous answer: Use a more sophisticated method (see next)
  - In practice: Judgement call, possibly based on business strategy
- The Bonferroni correction is a very simple method, but there are many other (**less conservative**) methods, including the closed testing procedure, the Boole-Bonferroni bound, and the Holm-Bonferroni method. This article on multiple comparisons contains more information, and this article contains more information about the false discovery rate (FDR), and methods for controlling that instead of the familywise error rate (FWER).

5.16. Analyzign Multiple Metrics

- Make sure/hope that multple metrics are moving towards the same direction (e.g., clicks versus stay time)

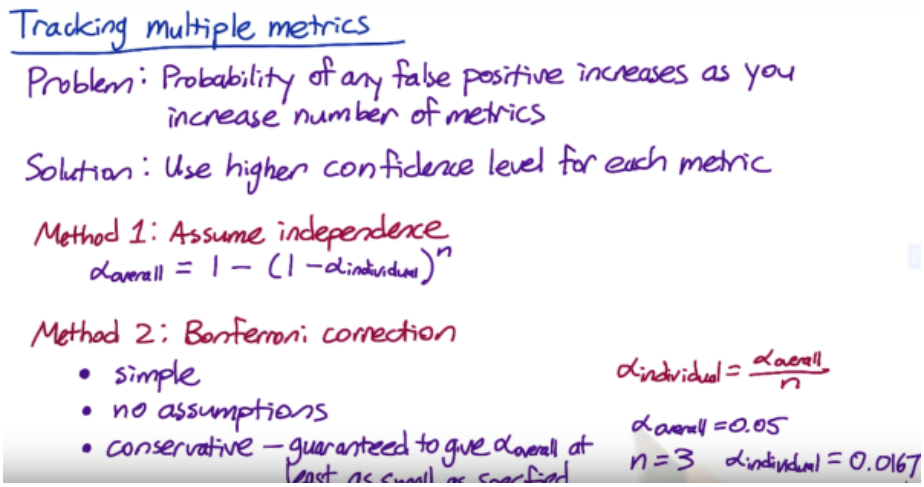


Figure 34: Tracking multiple metrics - Bonferroni

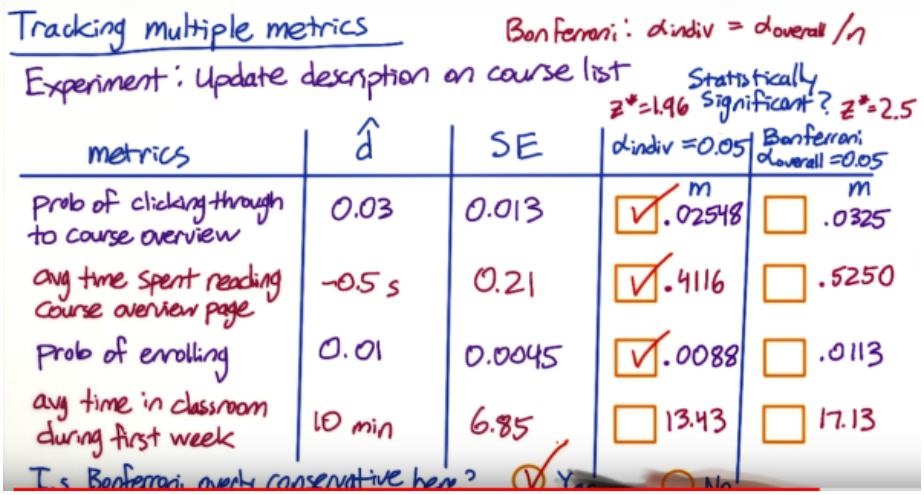


Figure 35: Tracking multiple metrics - Audacity

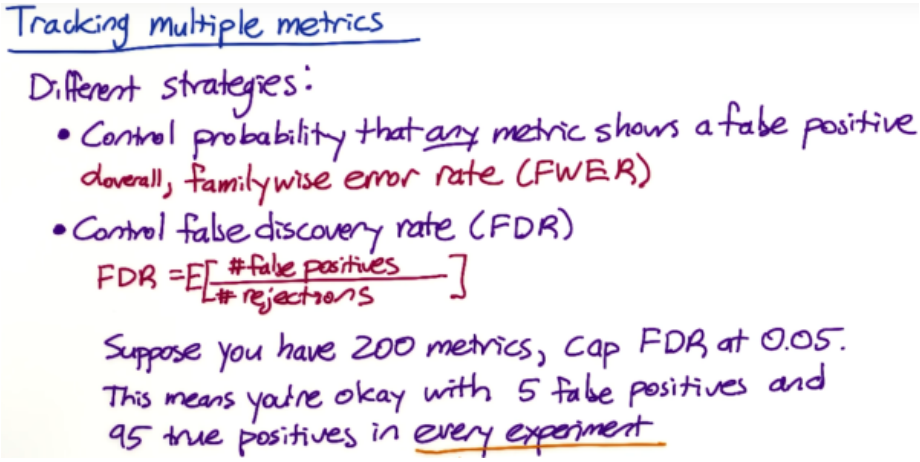


Figure 36: Tracking multiple metrics - FDR

- Better to come up with an OEC (Overall Evaluation Criterion) based on the business target of the company.

5.17. Draw Conclusions

- Examples of making recommendations can be found in Section 5.8 - 5.9 Single Metric and 5.12 - 5.15. Multiple Metrics
- Often need to go deeper to understand the user to find reasons for conflicting metrics or outcomes among cohorts.

5.18. Changes Over Time

- Ramp up the (sample size, user groups) of experiments. However, results may not be repeatable due to changes over time (i.e., due to seasonal-driven impact)
- We can keep a holdout/holdback group that don't get any chagnes to track seasonality effect.

Lesson 6: Final Project

As of March 1st, 2020, I have gone through the first five (video) lessons of the course, and will take an indefinite leave of absense from finishing the final project lesson. Some resources are listed as below.

- Project Instructions
- Final Project Template
- Project rubric

Reference

- Evan's Awesome A/B Tools
- Overlapping Experiment Infrastructure:More, Better, Faster Experimentation
- Large-Scale Validation and Analysis of Interleaved Search Evaluation