

Notes on A/B Testing (Udacity)

Yuanzhe Li

2020-02

Contents

Lesson 1: Overview of A/B Testing	3
1.15 Calculating confidence interval (CTR example)	3
1.17 Null and Alternative Hypothesis, <i>Two-tailed vs. One-tailed tests</i>	3
1.19 Pooled Standard Error	4
1.21 - 24. Sample Size and Power	4
1.25 Pooled Example	7
1.26 Confidence Interval Case Breakdown	7
Lesson 2: Policy and Ethics for Experiments	7
2.1 - 2.7. Four Principles	7
2.8 Accessing Data Sensitivity	7
2.10 Summary of Principles	9
Lesson 3: Choosing and Characterizing Metrics	9
3.2 - 3.3 Metric Definition Overview	9
3.5 Refining the Customer Funnel	9
3.6 - 3.7 Quizes on Choosing Metrics	10
3.8 Other techniques for defining metrics	11
3.10 - 11 Techniques to Gather Additional Data and Examples . .	11
3.13 Metric Definition: Click Through Example	11
3.16 - 3.17 Summary Metrics	14
3.18 - 3.19 Sensitivity and Robustness	14
3.20 Absolute Versus Relative Differences	15
3.21 - 3.22 Variability	15
3.24-25 Empirical Variability	15
Lesson 4: Designing an Experiment	17
4.2 - 4.3 Unit of Diversion Overview	18
4.4 - 4.5 Consistency of Diversion	19
4.6 - 4.7 Ethical Considerations	19
4.8 - 4.9 Unity of Analysis vs. Diversion	20

4.10 Inter- vs. Intra-User Experiments	21
4.11 - 4.13 Target Population, Cohort	21
4.16 - 4.18 Sizing Examples	22
4.20 - 22. Duration vs. Exposure	24
4.23 Learning Effects	24
Lesson 5: Analyzing Results	24
5.1 - 5.7 Sanity Checks (invariant metrics)	24
5.8 - 5.9 Single Metric	28
5.10 - 11. Simpson's Paradox	29
5.12 - 5.15. Multiple Metrics	32
5.16. Analyzign Multiple Metrics	32
5.17. Draw Conclusions	34
5.18. Changes Over Time	34
Lesson 6: Final Project	34
Reference	35

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

- Lesson 1: Overview of A/B Testing
 - 1.15 Calculating confidence interval (CTR example)
 - 1.17 Null and Alternative Hypothesis, *Two-tailed vs. One-tailed tests*
 - 1.19 Pooled Standard Error
 - 1.21 - 24. Sample Size and Power
 - 1.25 Pooled Example
 - 1.26 Confidence Interval Case Breakdown
- Lesson 2: Policy and Ethics for Experiments
 - 2.1 - 2.7. Four Principles
 - 2.8 Accessing Data Sensitivity
 - 2.10 Summary of Principles
- Lesson 3: Choosing and Characterizing Metrics
 - 3.2 - 3.3 Metric Definition Overview
 - 3.5 Refining the Customer Funnel
 - 3.6 - 3.7 Quizes on Choosing Metrics
 - 3.8 Other techniques for defining metrics
 - 3.10 - 11 Techniques to Gather Additional Data and Examples
 - 3.13 Metric Definition: Click Through Example
 - 3.16 - 3.17 Summary Metrics
 - 3.18 - 3.19 Sensitivity and Robustness
 - 3.20 Absolute Versus Relative Differences
 - 3.21 - 3.22 Variability
 - 3.24-25 Empirical Variability
- Lesson 4: Designing an Experiment
 - 4.2 - 4.3 Unit of Diversion Overview
 - 4.4 - 4.5 Consistency of Diversion

- 4.6 - 4.7 Ethical Considerations
- 4.8 - 4.9 Unity of Analysis vs. Diversion
- 4.10 Inter- vs. Intra-User Experiments
- 4.11 - 4.13 Target Population, Cohort
- 4.16 - 4.18 Sizing Examples
- 4.20 - 22. Duration vs. Exposure
- 4.23 Learning Effects
- Lesson 5: Analyzing Results
 - 5.1 - 5.7 Sanity Checks (invariant metrics)
 - 5.8 - 5.9 Single Metric
 - 5.10 - 11. Simpson's Paradox
 - 5.12 - 5.15. Multiple Metrics
 - 5.16. Analyzign Multiple Metrics
 - 5.17. Draw Conclusions
 - 5.18. Changes Over Time
- Lesson 6: Final Project
- Reference

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 successes.
- For $\alpha = 5\%$, we have $m = z_{0.025} \cdot \sqrt{\frac{0.1 \cdot 0.9}{1000}} = 0.019$, 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 $\hat{p}_{pool} = \frac{X_{cont} + X_{exp}}{N_{cont} + N_{exp}}$
- Pooled sample standard error $SE_{pool} = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{N_{cont}} + \frac{1}{N_{exp}}\right)}$
- 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 type 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:

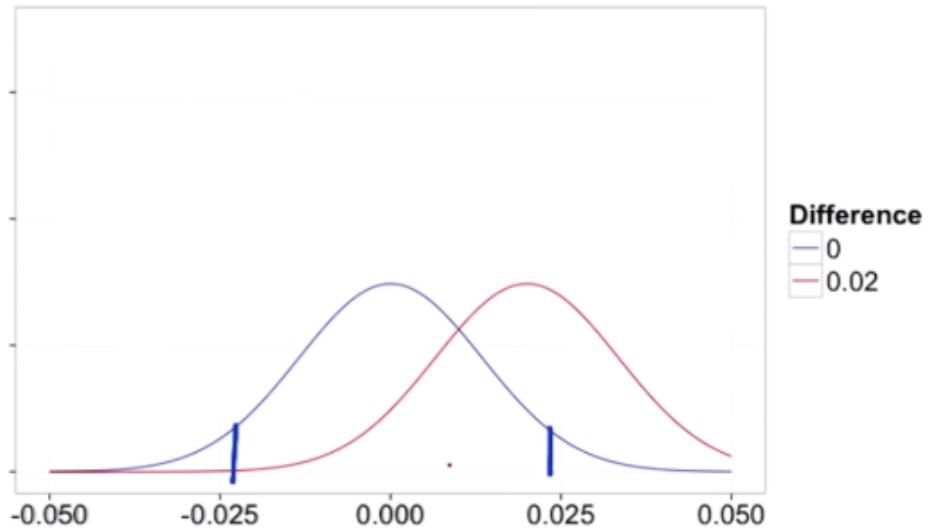


Figure 1: small_sample

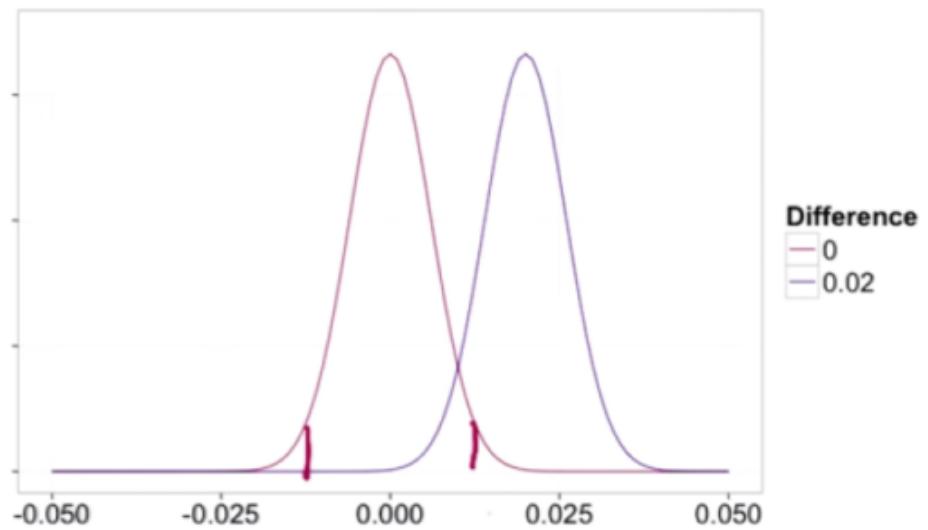


Figure 2: large_sample

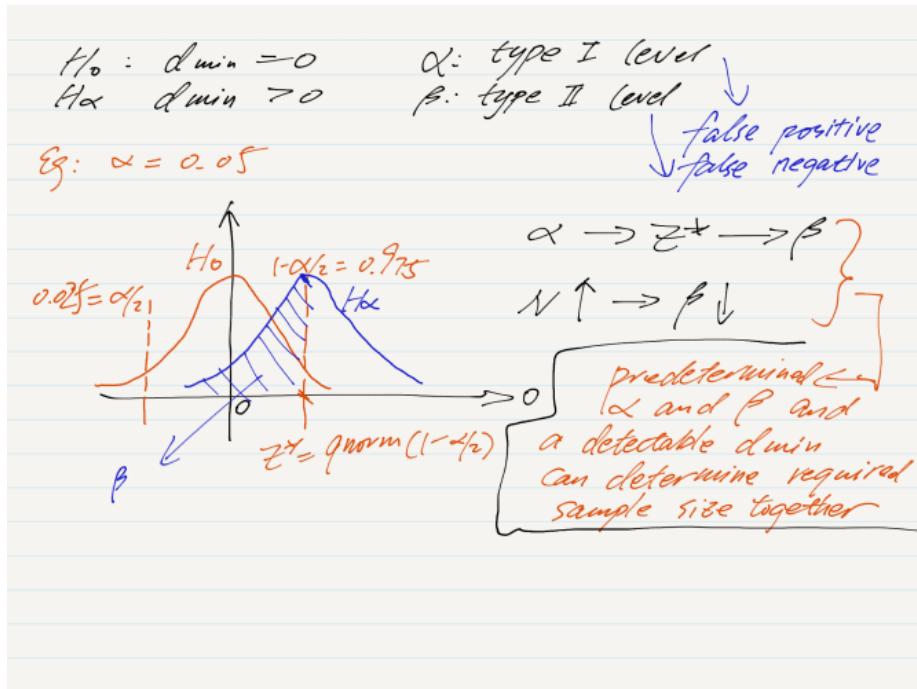


Figure 3: notes on required sample size

How number of page views varies

Change	Increase page views	Decrease page views
Higher click-through-probability in control (but still less than 0.5) $\sqrt{0.5 \times 0.5} = 0.5$ $\sqrt{0.1 \times 0.9} = 0.3$	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Increased practical significance level (d_{\min})	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Increased confidence level ($1 - \alpha$)	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Higher sensitivity ($1 - \beta$)	<input checked="" type="checkbox"/>	<input type="checkbox"/>

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 bound of the $1 - \alpha$ level CI $> d_{min} = 0.02$)

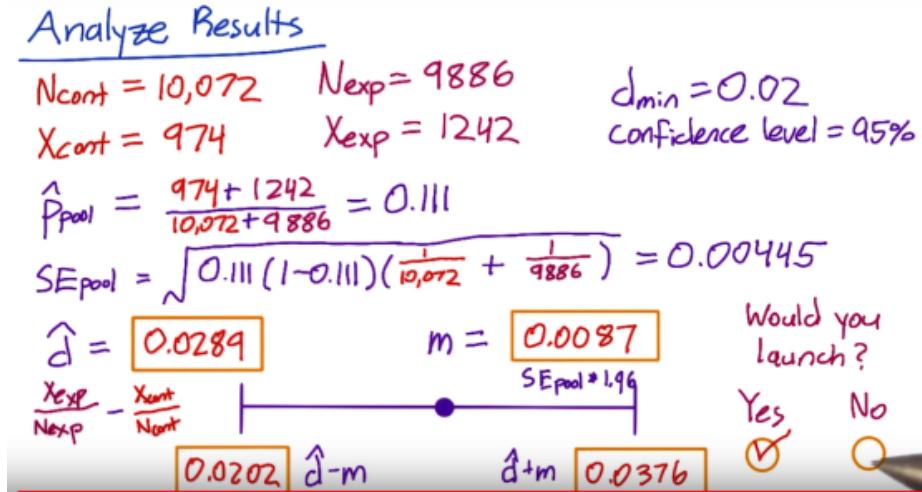


Figure 5: pooled example

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 experiments 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

Confidence Interval Cases

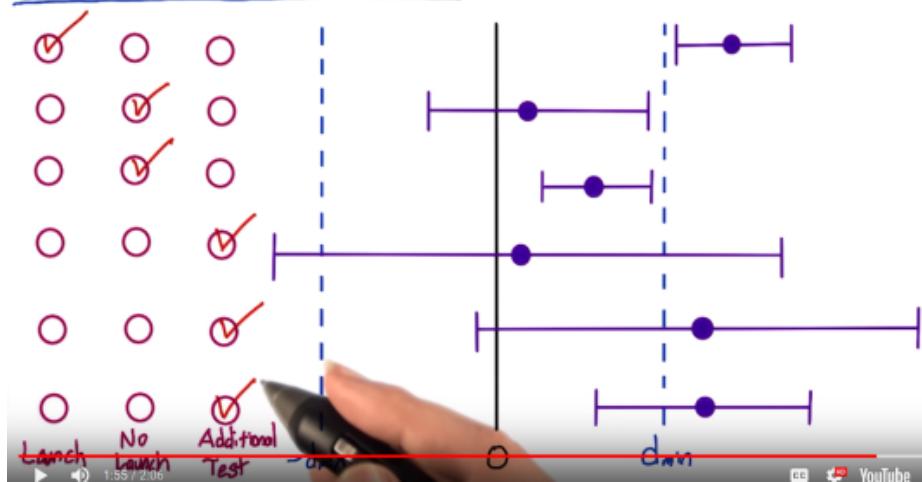


Figure 6: CI breakdown

Assessing Data Sensitivity

Should the following data be considered sensitive?

- | | |
|--|--|
| <input type="checkbox"/> Census data by zipcode | Too granular for re-identification. |
| <input type="checkbox"/> Daily traffic to specific sites | Can't identify individuals. |
| <input checked="" type="checkbox"/> Glucose levels with timestamps | Timestamps could identify. Regulation such as HIPAA. |
| <input type="checkbox"/> Online game stats | Game data not sensitive. |
| <input type="checkbox"/> Shopping stats by zipcode | Too granular for re-identification |
| <input checked="" type="checkbox"/> Credit card information | Very sensitive! |

Figure 7: assessing 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 multiple 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

Expanding on the funnel

- Homepage visits
- Exploring the site
 - # users who view course list
 - # users who view course details
- Create an account
 - # users who enroll in a course
 - # users who finish Lesson 1, lesson 2, etc
 - # users who sign up for coaching at various levels
- Completing a course
 - # users who enroll in a second class
 - # users who get jobs

- High-level metrics

Expanding on the funnel

Different platforms?

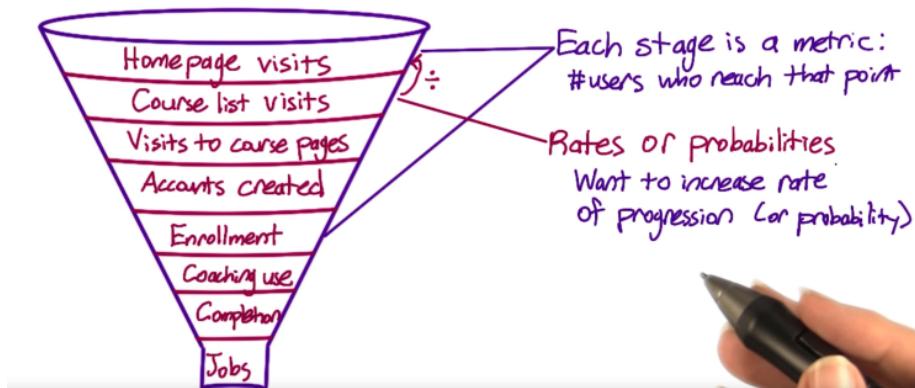


Figure 8: Metrics for funnels

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.,

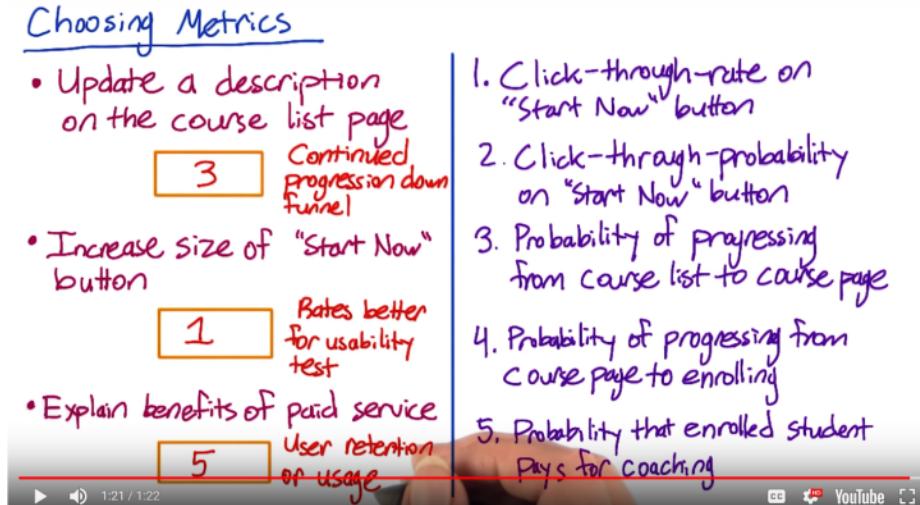


Figure 9: choosing metrics

- * 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

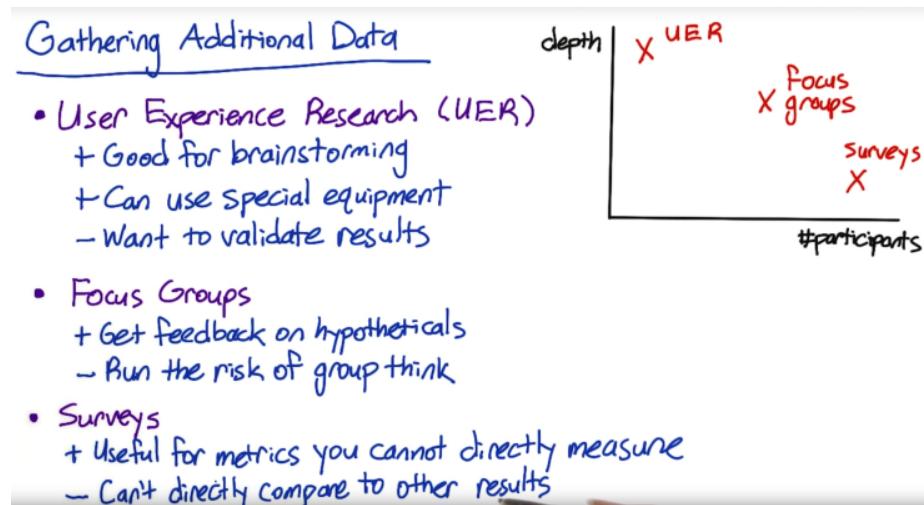


Figure 10: 3.10 techniques for getting additional data

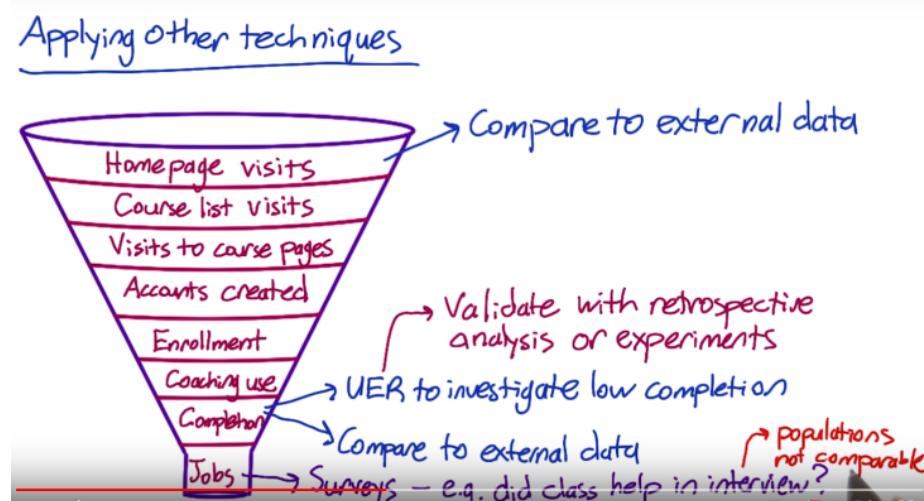


Figure 11: 3.11 gather data - udacity

Applying other techniques to difficult metrics

- Rate of returning for 2nd course
 - Survey → proxy
 - Average happiness of shoppers
 - Survey
 - UER
 - Probability of finding information via search
 - External data
 - Human evaluation
- UFR
- Possible proxies:
• Time spent
• Clicks on results
• Follow-up queries

Figure 12: 3.12 when there is no data

Defining a metric

High-level metric: Click-through-probability = $\frac{\# \text{ users who click}}{\# \text{ users who visit}}$

Def #1: For each <time interval>, $\frac{\# \text{ cookies that click}}{\# \text{ cookies}}$

Def #2: $\frac{\# \text{ pageviews w/ click}}{\# \text{ pageviews}}$ within <time interval>

Def #3: $\frac{\# \text{ clicks}}{\# \text{ pageviews}}$ (click-through-rate)

Which metrics have which problems?

	1: Cookie Prob	2: Pageview Prob	3: Rate
Double click	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Back button caches page	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Click-tracking bug	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



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)	$\text{Var}(X) + \text{Var}(Y)$
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 $N = [87029, 113407, 84843, 104994, 99327, 92052, 60684]$
 - $\bar{N} = 92,052$, $\hat{\sigma} = 17,015$
 - $SE = \frac{\hat{\sigma}}{\sqrt{N}} = 6,430$
 - Margin of error $m = z^* \cdot SE = 1.96 \cdot 6,430 = 12,605$
 - The CI is 79,158 to 104,367

3.24-25 Empirical Variability

- Uses of A/A tests

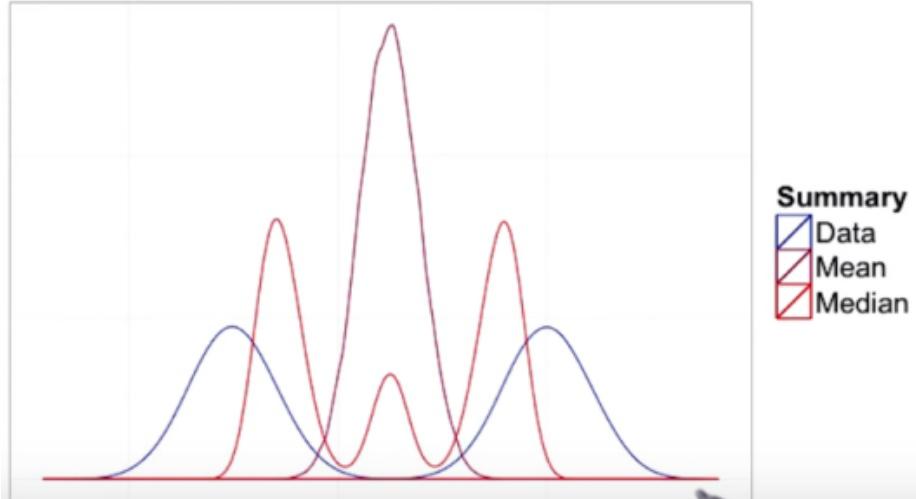


Figure 13: bimodal distribution of latency

- 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:

$$m = SD \cdot z^*$$

$$= 0.059 \cdot 1.96 = 0.116 \text{ empirically}$$

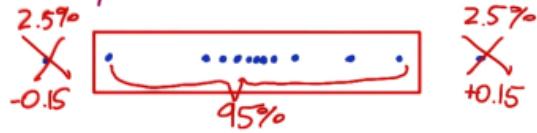
$$\text{Analytically: } SE = \sqrt{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}}) \left(\frac{1}{N_{\text{cont}}} + \frac{1}{N_{\text{exp}}}\right)}$$

Slightly different margin of error for each experiment

- Directly estimate confidence interval from empirical distribution

Calculating variability empirically

Directly estimate confidence interval:



Since we have 20 data points, dropping the highest and the lowest gives a 90% confidence level: -0.1 to 0.06

Empirical standard deviation: $0.059 * 1.65 = 0.097$
z-score for 90% confidence



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

Lesson 4: Designing an Experiment

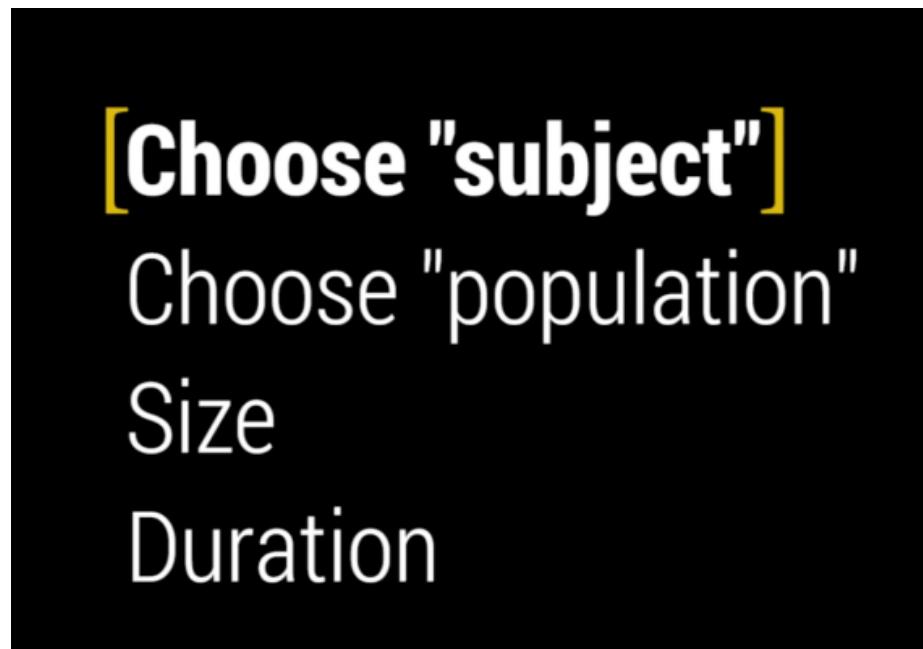


Figure 14: Outline of lesson 4

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

		desktop homepage	sign in	visit class	watch video	mobile auto sign in	watch video
user-id		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
cookie		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
event		<input checked="" type="checkbox"/>					
device id		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
IP address		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

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 experiment. If learning effect is detected later, we can switch to a cookie-based experiment.
- Example

Which unit of diversion will give enough consistency?

Experiment	Event	Cookie	User-id
Change reducing video load time Users probably won't notice	✓	○	○
Change button color and size Distracting if button changes on reload Different look on different devices ok	○	✓	○
Change order of search results Users probably won't notice	✓	○	○
Add Instructor's Notes before quizzes Users will almost certainly notice Cross-device consistency important	○	○	✓

Figure 16: Example of consistency needed from unit of diversion

4.6 - 4.7 Ethical Considerations

- An example is as follows.
 - Notice that only the second case requires additional ethical review/consent from the user because it might compromise the anonymity of cookie-based data.

Ethical considerations

Which experiments might require additional ethical review?

- Newsletter prompt after starting course User id diversion
 - No new information being collected
 - Fine if original data collection was approved
- Newsletter prompt on course overview Cookie diversion
 - Depends: Are email addresses stored by cookie?
 - Potentially impacts other data collection
- Changes course overview page Cookie diversion
 - Not a problem, and probably already being done

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.

Unit of analysis and unit of diversion

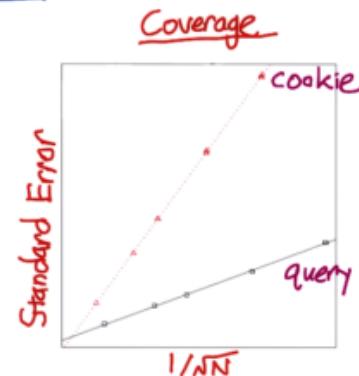
Measure variability of a metric

Unit of diversion: query or cookie

Metric: Coverage = $\frac{\text{#queries with ad}}{\text{#queries}}$

Unit of analysis: query

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



When unit of analysis = unit of diversion,
variability tends to be lower and closer to analytical estimate

Figure 17: Unit of analysis versus diversion example

- The quiz example.

Unit of analysis and unit of diversion

When would you expect the analytic variance to match the empirical variance?

- Metric: click-through-rate = $\frac{\# \text{clicks}}{\# \text{pageviews}}$ Unit of analysis: pageview
Unit of diversion: cookie
- Metric: #cookies that view homepage Unit of analysis: cookie
Unit of diversion: ~~pageview~~ cookie user-id "larger" than unit of diversion!
Metric not well-defined
- Metric: $\frac{\# \text{users who sign up for coaching}}{\# \text{users enrolled in any course}}$ Unit of analysis: user-id
Unit of diversion: user-id

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, \dots, X_N in that order, and algorithm Y would show Y_1, Y_2, \dots, Y_N . An interleaved experiment would show some interleaving of those results, for example, $X_1, Y_1, X_2, Y_2, \dots$ 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](#).

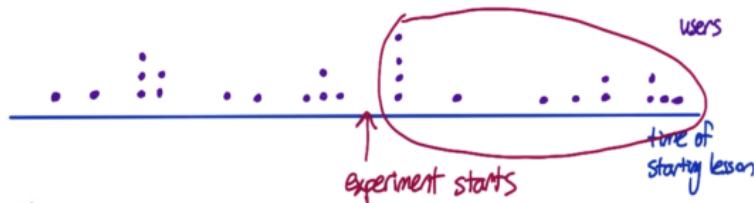
4.11 - 4.13 Target Population, Cohort

- 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:

Using cohorts in experiments

Audacity example: Have existing course and change structure of lesson

Unit of diversion: user-id - but, can't run on all users in course



Control: Needs to be a comparable cohort

Cohorts limit your experiment to a subset of the population - can affect variability

4.16 - 4.18 Sizing Examples

- How variability affects sizing

How variability affects sizing

Audacity includes promotions for coaching next to videos

Experiment: Change wording of message

Metric: Click-through rate = $\frac{\# \text{clicks}}{\# \text{pageviews}}$

Unit of diversion: Pageview, or cookie

Analytic variability won't change, but probably under-estimate for cookie diversion

Empirical estimate with 5000 pageviews

By pageview: 0.00515

By cookie: 0.0119

Figure 19: sizing example 1

- See the [codes/empirical-sizing.r](#) for code example.
- Quiz

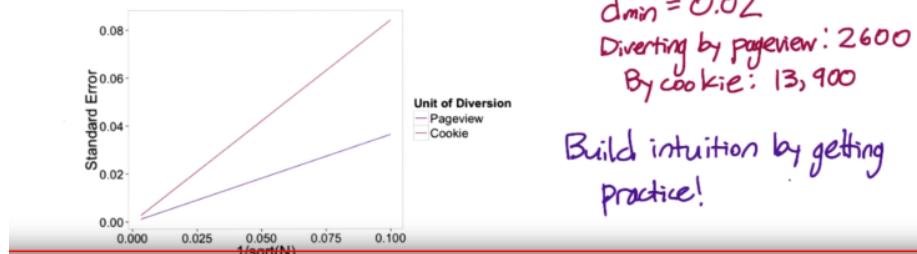
How variability affects sizing

Empirical estimate with 5000 pageviews

By pageview: 0.00515

By cookie: 0.0119

To calculate size, assume $SE \sim \frac{1}{\sqrt{N}}$



$$d_{min} = 0.02$$

Diverting by pageview: 2600

By cookie: 13,900

Build intuition by getting practice!

Figure 20: sizing example 2

How to reduce the size of an experiment

Experiment: Change order of courses on course list

Metric: Click-through-rate

$$d = 0.05 \quad \beta = 0.2$$

Unit-of-diversion: cookie

$$d_{min} = 0.01 \quad SE = 0.0628$$

for 600 pageviews

Result: Need 300,000 pageviews per group!

Which strategies could reduce the number of pageviews?

Increase d_{min} , d , or β

Change unit of diversion to page view

Target experiment to specific traffic

Change metric to cookie-based click-through-probability

Figure 21: sizing quiz 1

How to reduce the size of an experiment

- Change unit of diversion to page view
 - Makes unit of diversion same as unit of analysis
 - But will less consistent experience be okay?
 - If SE changes to 0.0209 → only 34,000 pageviews per group
- Target experiment to specific traffic
 - Non-English traffic will dilute the results
 - Could impact choice of practical significance boundary
 - SE changes to 0.0188, down to 0.015 → only 12,000 pageviews per group
- Change metric to cookie-based click-through-probability
 - Often doesn't make significant difference
 - If there is a difference, variability would probably go down

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. novelty 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

Duration vs Exposure

Size of an experiment : 1 million pageviews

Average traffic per day : 500,000 page views

Run experiment for 2 days

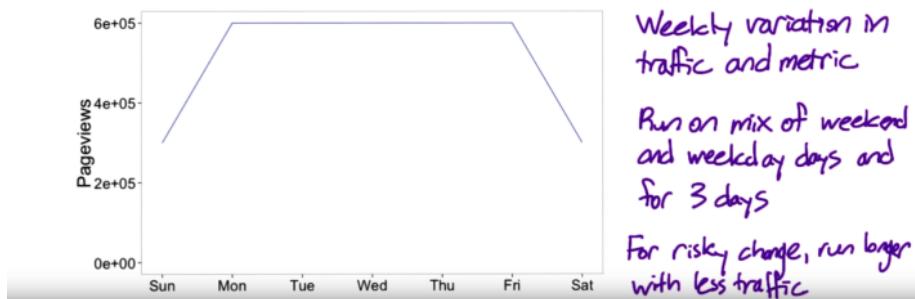


Figure 23: Duration vs exposure: example

When to limit exposure

Which experiments are risky enough that Audacity might want to limit the number of users exposed?

- Changes database . If this goes wrong, effects could be huge!
- Changes color of "Start Now" button Low risk (but should still test)
- Allows Facebook login If you don't roll out, how to deal with Facebook logins?
- Changes order of courses on course list Low risk if you've run similar experiments

Figure 24: When to limit exposure

- Check *population sizing metrics* to make sure control and experiment groups are comparable
- Check actual invariant metrics
- Quizzes
 - population sizing 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	<input checked="" type="checkbox"/> random	<input checked="" type="checkbox"/> not directly but should be split evenly	<input checked="" type="checkbox"/> randomized	<input checked="" type="checkbox"/> happens before course list	<input type="checkbox"/> could be affected
Changes infrastructure to reduce load time Unit of diversion: event	<input checked="" type="checkbox"/> "larger" than	<input checked="" type="checkbox"/> unit of diversion	<input checked="" type="checkbox"/> random	<input checked="" type="checkbox"/> happens before viewing videos	<input type="checkbox"/> can't be tracked

Figure 25: Quiz - 5.3 choosing population sizing metrics

- invariant metrics

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?

- | | |
|---|--|
| <input checked="" type="checkbox"/> # events | This, #cookies, and #users all good |
| <input type="checkbox"/> CTR on "Start Now" | Adding sign-in button to home page could affect this |
| <input type="checkbox"/> Probability of enrolling | Users often enroll after signing in |
| <input type="checkbox"/> Sign-in rate | This is what we're trying to change! |
| <input checked="" type="checkbox"/> Video load time | No backend changes |

Figure 26: Quiz - 5.4 invariant metrics

- Checking invariants

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

Write the steps you would take in plain English.

Then discuss on the forums and check this box:

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

$$0.5 \text{ of success } SD = \sqrt{\frac{0.5 \cdot 0.5}{64,454 + 61,818}} = 0.0014$$

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

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	2439	0.537
Sun	3476	3025	0.532

Checking invariants

Run experiment for 2 weeks.
Unit of diversion: cookie

Total control: 64,454

Total experiment: 61,818

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

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

Analysis with a single metric

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

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

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

	control clicks	control pageviews	experiment clicks	experiment pageviews	Sanity check: pass
Day 1	51	1292	115	1305	Empirical SE: 0.0035 w/ 10,000
Day 2	39	853	73	835	Pageviews per group
Day 3	64	1129	91	1133	$SE \sim \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Day 4	43	873	60	871	$\frac{0.0035}{\sqrt{\frac{1}{10000} + \frac{1}{10000}}} = \frac{SE}{\sqrt{\frac{1}{10000} + \frac{1}{10000}}}$
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

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

$$d = 0.05 \quad \beta = 0.2$$

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

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

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

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

$$m = 0.0041 * 1.96 = 0.0080$$

Confidence interval: 0.0020 to 0.0380

Recommendation: Launch

Sanity check: pass

Empirical SE:

$$0.0035 \text{ w/ 10,000}$$

Pageviews per group

$$SE \sim \sqrt{\frac{1}{N_{cont}} + \frac{1}{N_{exp}}}$$

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

$$SE = 0.0041$$

Figure 28: Quiz 5.9 - 2

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

- Test the probability of observing 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.

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 $\alpha = 0.05 \quad \beta = 0.2$

	control clicks (CTR)	control pageviews	experiment clicks (CTR)	experiment pageviews	
Day 1	51 (.031)	1242	115 (.088)	1305	Sanity check: pass
Day 2	39 (.046)	853	73 (.087)	835	# days: 7
Day 3	64 (.057)	1129	91 (.090)	1133	# days with positive change: 7
Day 4	43 (.049)	873	60 (.061)	871	If no difference, 50% chance of positive change on each day
Day 5	55 (.046)	1197	78 (.061)	1134	
Day 6	44 (.043)	1023	72 (.071)	1015	
Day 7	56 (.056)	1003	76 (.078)	977	
Total	352 (.048)	7370	565 (.078)	7270	Cannot assume normal

Figure 29: Quiz 5.9 - 3

Analysis with a single metric

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

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

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

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

$$m = 0.0026 * 1.96 = 0.0051 \quad \text{Confidence Interval: } 0.0065 \text{ to } 0.0167$$

Figure 30: Quiz 5.9 - 4

Analysis with a single metric

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

Days where CTR is higher in experiment: 9 / 14

Week 1				Week 2					
	Control CTR	Control #pageviews	Experiment CTR	Experiment #pageviews		Control CTR	Control #pageviews	Experiment CTR	Experiment #pageviews
Mon	0.097	1929	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	2012
Thu	0.109	1985	0.087	2015	✓ Thu	0.097	1977	0.101	2023
Fri	0.107	1973	0.094	2027	✓ Fri	0.097	2019	0.101	1981
✓ Sat	0.092	2021	0.147	1979	✓ Sat	0.110	2035	0.151	1965
✓ Sun	0.110	2041	0.142	1959	✓ Sun	0.096	2007	0.150	1993

Figure 31: Quiz 5.9 - 5

Analysis with a single metric

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

Weedays: -0.0078
to 0.0043

Weekends: 0.0361
to 0.0553

Effect size



0.0065

0.0167

Statistically significant?

- Yes No

Sign test

p-value: 0.4240

Statistically significant?

- Yes No

Recommendation: Do not launch (yet).

Figure 32: Quiz 5.9 - 6

Simpson's paradox Recommendation: Dig deeper

	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)

Wait — why are there more pageviews from new users in the control group?!

- Something wrong with set-up
- Change affects new users and experienced users differently

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)

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

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

Method 2: Bonferroni correction

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

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

$$\begin{aligned} \alpha_{\text{overall}} &= 0.05 \\ n &= 3 \quad \alpha_{\text{individual}} = 0.0167 \end{aligned}$$

Figure 34: Tracking multiple metrics - Bonferroni

Tracking multiple metrics

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

Experiment: Update description on course list Statistically significant? $Z^* = 2.5$

metrics	\hat{d}	SE	$\alpha_{\text{indiv}} = 0.05$	Bonferroni: $\alpha_{\text{overall}} = 0.05$
Prob of clicking through to course overview	0.03	0.013	<input checked="" type="checkbox"/> m	<input type="checkbox"/> .0325
Avg time spent reading course overview page	-0.5 s	0.21	<input checked="" type="checkbox"/> .4116	<input type="checkbox"/> .5250
Prob of enrolling	0.01	0.0045	<input checked="" type="checkbox"/> .0088	<input type="checkbox"/> .0113
Avg time in classroom during first week	10 min	6.85	<input type="checkbox"/> 13.43	<input type="checkbox"/> 17.13
I: Bonferroni: overly conservative here?			<input checked="" type="checkbox"/> Y	<input type="checkbox"/> No

Figure 35: Tracking multiple metrics - Audacity

Tracking multiple metrics

Different strategies:

- Control probability that any metric shows a false positive overall, familywise error rate (FWER)
- Control false discovery rate (FDR)
$$FDR = E\left[\frac{\# \text{false positives}}{\# \text{rejections}}\right]$$

Suppose you have 200 metrics, cap FDR at 0.05.

This means you're okay with 5 false positives and 95 true positives in every experiment.

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 changes 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 absence 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