A/B testing

**2. Setting the target metrics**

Click through rate

Number of clicks on the target of test ÷ number of page views

Click through probability

Number of unique users having clicked on the target of test ÷ number of users having viewed the page

**Difference**

**Calculation of click through rate and click through probability**

Suppose there are two users visiting a page, and the target is to see how different it is if a button in that page is green or red. Two users visited the page, and A viewed the page and left, didn't click on the button. B viewed the page and clicked the button for 5 times. Then the click through rate is 5÷2=2.5. The click through probability is 1÷2=0.5.

**When to use these two measurements**

Click through rate is useful when we want to see the impact of usability, i.e. with a change in the page, how likely is it that a user will notice and click the button. Click through probability is helpful when we want to measure the total impact, because it doesn't matter if a user double clicked or the page has a reload or any technical issues.

**3. Pooled standard error**

Control group and experiment group:

Number of users who clicked in each group: X control and X experiment

Total number of users in each group: N control and N experiment

Pooled probability of click: the total number of users who clicked ÷ the total number of users

= (X control + X experiment) ÷ (N control + N experiment)

Standard error pooled probability SE pool =

The target is to estimate the difference d between P experiment and P control.

H0: d = 0 and d ~ N(0, SE pool)

If d >= 1.96 \* SE pool or d <= -1.96 \* SE pool, then reject H0 and in favor of the alternative.

**4. Policy and ethical considerations when designing an A/B test**

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. The harm considered encompasses physical, psychological and emotional, social, and economic concerns. If the risk exceeds minimal risk, then informed consent is required.

Benefits: *what benefits might result from the study*?

Even if the risk is minimal, how might the results help? In most online A/B testing, the benefits are around improving the product. In other social sciences, it is about understanding the human condition in ways that might help, for example in education and development. In medicine, the risks are often higher but the benefits are often around improved health outcomes. It is important to be able to state what the benefit would be from completing the study.

Choice: *what other choices do participants have*?

For example, if you are testing out changes to a search engine, participants always have the choice to use another search engine. The main issue is that the fewer alternatives that participants have, the more issue that there is around coercion and whether participants really have a choice in whether to participate or not, and how that balances against the risks and benefits.

For example, in medical clinical trials testing out new drugs for cancer, given that the other main choice that most participants face is death, the risk allowable for participants, given informed consent, is quite high.

In online experiments, the issues to consider are what the other alternative services that a user might have, and what the switching costs might be, in terms of time, money, information, etc.

Privacy: *what data is being collected, and what is the expectation of privacy and confidentiality*?

This last question is quite nuanced, encompassing numerous questions:

* Do participants understand what data is being collected about them?
* What harm would befall them should that data be made public?
* Would they expect that data to be considered private and confidential?

For example, if participants are being observed in a public setting (e.g., a football stadium), there is really no expectation of privacy. If the study is on existing public data, then there is also no expectation of further confidentiality.

If, however, new data is being gathered, then the questions come down to:

* What data is being gathered? How sensitive is it? Does it include financial and health data?
* Can the data being gathered be tied to the individual, i.e., is it considered personally identifiable?
* How is the data being handled, with what security? What level of confidentiality can participants expect?
* What harm would befall the individual should the data become public, where the harm would encompass health, psychological / emotional, social, and financial concerns?

For example, often times, collected data from observed “public” behavior, surveys, and interviews, if the data were not personally identifiable, would be considered exempt from IRB review (reference: NSF FAQ below).

To summarize, there are really three main issues with data collection with regards to experiments:

* For new data being collected and stored, how sensitive is the data and what are the internal safeguards for handling that data? E.g., what access controls are there, how are breaches to that security caught and managed, etc.?
* Then, for that data, how will it be used and how will participants’ data be protected? How are participants guaranteed that their data, which was collected for use in the study, will not be used for some other purpose? This becomes more important as the sensitivity of the data increases.
* Finally, what data may be published more broadly, and does that introduce any additional risk to the participants?

**Difference between pseudonymous and anonymous data**

One question that frequently gets asked is what the difference is between identified, pseudonymous, and anonymous data is.

**Identified** data means that data is stored and collected with personally identifiable information. This can be names, IDs such as a social security number or driver’s license ID, phone numbers, etc. HIPAA is a common standard, and that standard has [18 identifiers (see the Safe Harbor method)](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#standard) that it considers personally identifiable. Device id, such as a smartphone’s device id, are considered personally identifiable in many instances.

**Anonymous** data means that data is stored and collected without any personally identifiable information. This data can be considered **pseudonymous** if it is stored with a randomly generated id such as a cookie that gets assigned on some event, such as the first time that a user goes to an app or website and does not have such an id stored.

In most cases, anonymous data still has time-stamps -- which is one of the HIPAA 18 identifiers. Why? Well, we need to distinguish between anonymous data and anonymized data. **Anonymized data** is identified or anonymous data that has been looked at and guaranteed in some way that the re-identification risk is low to non-existent, i.e., that given the data, it would be hard to impossible for someone to be able to figure out which individual this data refers to. Often times, this guarantee is done statistically, and looks at how many individuals would fall into every possible bucket (i.e., combination of values).

What this means is that anonymous data may still have high re-identification risk (see [AOL example](http://en.wikipedia.org/wiki/AOL_search_data_leak)).

So, if we go back to the data being gathered, collected, stored, and used in the experiment, the questions are:

* How sensitive is the data?
* What is the re-identification risk of individuals from the data?

As the sensitivity and the risk increases, then the level of data protection must increase: confidentiality, access control, security, monitoring & auditing, etc.

**Additional reading**

Daniel Solove's ["A Taxonomy of Privacy"](https://www.law.upenn.edu/journals/lawreview/articles/volume154/issue3/Solove154U.Pa.L.Rev.477%282006%29.pdf) classifies some of things people mean by privacy in order to better understand privacy violations.

**Summary of principles**

It is a grey area as to whether many of these Internet studies should be subject to IRB review or not and whether informed consent is required. Neither has been common to date.

Most studies, due to the nature of the online service, are likely minimal risk, and the bigger question is about data collection with regards to identifiability, privacy, and confidentiality / security. That said, arguably, a neutral third party outside of the company should be making these calls rather than someone with a vested interest in the outcome. One growing risk in online studies is that of bias and the potential for discrimination, such as differential pricing and whether that is discriminatory to a particular population for example. Discussing those types of biases is beyond the scope of this course.

Our recommendation is that there should be internal reviews of all proposed studies by experts regarding the questions:

* Are participants facing more than minimal risk?
* Do participants understand what data is being gathered?
* Is that data identifiable?
* How is the data handled?

And if enough flags are raised, that an external review happen.

**Internal process recommendations**

Finally, regarding internal process of data handling, we recommend that:

1. Every employee who might be involved in A/B test be educated about the ethics and the protection of the participants. Clearly there are other areas of ethics beyond what we’ve covered that discuss integrity, competence, and responsibility, but those generally are broader than protecting participants of A/B tests (cite ACM code of ethics).
2. All data, identified or not, be stored securely, with access limited to those who need it to complete their job. Access should be time limited. There should be clear policies of what data usages are acceptable and not acceptable. Moreover, all usage of the data should be logged and audited regularly for violations.
3. You create a clear escalation path for how to handle cases where there is even possibly more than minimal risk or data sensitivity issues.

**Additional reading:**

* [Belmont Report](http://www.hhs.gov/ohrp/policy/belmont.html)
* [Common Rule definition](http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#subparta)
* [NSF guidelines](http://www.nsf.gov/bfa/dias/policy/human.jsp)
* [NSF FAQ for Social Science & Behavioural research](http://www.nsf.gov/bfa/dias/policy/hsfaqs.jsp#exempt)
* [HHS IRB Guidebook](http://www.hhs.gov/ohrp/archive/irb/irb_chapter1.htm)
  + [Definition of Minimal Risk](http://www.hhs.gov/ohrp/archive/irb/irb_chapter3.htm)
  + [Discussion of different types of data gathering](http://www.hhs.gov/ohrp/archive/irb/irb_chapter4.htm)
* [UTexas overview](http://www.utexas.edu/research/rsc/humansubjects/whatis.html)
* [UC Irvine overview](http://www.research.uci.edu/compliance/human-research-protections/researchers/activities-irb-review.html)
* The Association for Computer Machinery has developed a [code of ethics](http://www.acm.org/about/code-of-ethics).
* As an example, there’s a thorough outline of an [“ideal” ethical privacy design](http://www.oii.ox.ac.uk/research/projects/?id=107) for mobile connectivity measurements that could be used as a model.

**Design metrics**

1. Time consideration: should not take too long

2. Ease of access to data: how easy is it to access the data

**Techniques to gather additional data**

1. User experience research

+ Good for brainstorming

+ Can use special equipment, e.g. eye tracking

- Want to validate results, sample size is too small

2. Focus groups

+ Get feedback on hypotheticals

- Run the risk of group thinking

3. Surveys

+ Useful for metrics you cannot directly measure

- Cannot directly compare to other results, participants might not be telling the truth

**Uses of A/A tests**

1. Compare resutls to what you expect (sanity check)

Example

* 20 experiments, each on 0.5% of the traffic - 50 users in each group
* 20 more experiments, each on 1% of the traffic - 100 users in each group
* 10 more experiments, each on 5% of the traffic - 500 users in each group
* How many experiments will show a statistically significant difference at the 95% confidence level?

2. Estimate variance and calculate confidence

3. Directly estimate confidence interval

**Unit of diversion**

1. User ID

* Stable and unchanging. When designing an A/B test, a user based on the user ID can be assigned either to the control group or the experiment group.
* Personally identifiable. Normally associated with other personal information.

2. Anonymous ID

* Cookie assigned to a user when visiting the website. Specific to a browser or a device. Changing device will get a different cookie.
* Users can clear cookies.

3. Event

* No consistent experiment
* Use only for non-user visible changes
* Often used for e.g. changing the rank of the items

4. Device ID

* Only available for mobile devices
* Tied to specific device, unchangeable by a user

5. IP address

* Changes when location changes

Choice of unit of diversion

* User ID user experience keeps the same as long as they are signed in. Thus for testing features in the signed in part user ID can be used.
* Cookies are used when testing features that cross the sign-in sign-out border. E.g. change the location of the sign-in bar. Using cookies can guarantee the user experience to be consistant across the sign-in and sign-out border, but not across devices.

Examples

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment** | **Event** | **Cookie** | **User ID** | **Reason** |
| Change reducing video load time | X |  |  | Users probably won't notice if the page loading is slower or faster. |
| Change button color and size |  | X |  | Can be distracting if the button changes when a user reloads the page. But it is fine if the UI looks different on different devices. |
| Change order of search results | X |  |  | Users probably won't notice. |
| Add instructor notes before quizzes |  |  | X | Cross device consistency is very import in this case. |

Unit of diversion

**Intra user experiment**

Expose the same user to a particular feature on and off in a user time period.

-- Need to be careful to choose a comparable time window. E.g. not before two weeks of a holiday so people behave really different.

-- On and off learning problem. It takes some time for a user to learn certain feature, and after some time when the feature disappears user gets influenced thus get different user behaviors.

**Cohort**

Define an entering class and only look at users who entered the experiments on both sides around the same time, and go forward from there.

Use cohort instead of population when:

Looking for learning effects

Examining user retention

Want to increase user activity

Anything require users to be established

E.g. look at usage of a site or a particular mobile device.

**Empirical variability vs. analytical variability**

Sometimes the empirically computed variability is much higher than the analytically computed variability. This happens when the unit of analysis is different than the unit of diversion.

The unit of analysis is basically the denominator of the metric. So for example, if you're doing click through rate, and you have clicks divided by page views then page view would be your unit of analysis. When the unit of diversion is also a page view, so as would be the case in an event base diversion, then the analytically computed variability is likely to be very close to the empirically computed variability. If, however, the unit of diversion is a cookie or a user id then the variability of the same metric click through rate is actually going to be much higher. Sometimes by a factor of four, five, maybe even more. And in those cases you really want to move to an empirically computed variability given your unit of diversion. The reason for such big difference is when you're actually computing variability analytically, you're fundamentally making an assumption about the distribution of the data. But you're not just making an assumption about the distribution of the data. You're also making an assumption about what's considered to be independent. You're basically doing these random draws of whether they're independent or not. When you're doing event-based diversion every single event is a different random draw, and so your independence assumption is actually valid. Now when you're doing cookie or user ID based diversion, that independence assumption is no longer valid because you're actually diverting groups of events. So they're actually correlated together. That will increase your variability greatly.

**How to reduce the size of an experiment**

Experiment: change order of courses on course list

Metric: click-through-rate

Unit of diversion: cookie

Parameters:

* alpha = 0.05
* beta = 0.2
* d-min = 0.01
* SE = 0.0628
* for 1000 page views

Result: need 300000 page views per group

**What strategies could reduce the number of page views?**

1. Increase d-min, alpha, or beta

2. Change unit of diversion to page view

3. Target experiment to specific traffic

How to evaluate A/B testing results

1. Sanity check

2. Single values

3. Multiple values

4. Gotchas

**Checking invariants**

**Example scenario**

Run experiments for 2 weeks and unit of diversion is cookie.

Total control: 64454

Total experiment: 61818

Given each cookie is randomly assigned to the control experiment group with probability of 0.5, how to figure out whether the difference is within expectations?

Suppose a user can be randomly put into either control or experiment group (similar with tossing a coin), then it follows a binomial distribution, thus the question is

1. Compute standard error of binomial with probability 0.5 of success

SE = sqrt (0.5 \* 0.5 / (64454 + 61818)) = 0.001407

2. Multiple by z-score to get margin of error

M = SE \* 1.96 (suppose the significance level is 0.05) = 0.0027

3. Compute confidence interval around 0.5

Lower bound = 0.5 - 0.0027 = 0.4973

Higher bound = 0.5 + 0.0027 = 0.5027

4. Check whether observed fraction is within interval

Observed fraction = 64454 / (64454 + 61818) = 0.5104

This is significantly greater than the higher bound, thus the difference is beyond expectations.

What to do in case there is something going wrong with the experiment setup:

* 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

If sanity check fails, don't proceed with any subsequent steps for A/B testing, instead, try to understand why sanity checks fail. Measures can be e.g. understand from engineers if the experiment setup is correct.

Learning effect can be a factor influencing the experiment, but normally it shows more significant influence as time goes. If some big changes show at the beginning, then it should not be because of learning effect.

**Analysis with single variant**

**Example scenario**

Machine generated alternative text:
DAT 2 
DOI 22 
DOT 5 
CliGk 
51 con) 
(.09') 
6H 1.057) 
1.43 C. 
55 (.0%) 
(.0H3) 
5b (.05') 
1242 
1129 
373 
1023 
1003 
1370 
chas 
115 
73 ( .087) 
ql (OSO) 
60 C.oq) 
73 (.oq) 
72 
7b (.03) 
5 51.01 
335 
1133 
371 
119t 
1015 
q 77 

Experiment: change color and placement of "start now" button.

Metric: click through rate

Unit of diversion: cookie

Practical significance boundary: d-min = 0.01

Alpha = 0.05

Beta = 0.02

X-cont = 352

N-cont = 7370

X-exp = 565

N-exp = 7270

Empirical SE: 0.0035 with 10000 page views per group

SE ~ sqrt(1/N1 + 1/N2)

0.0035 / sqrt(1/10000 + 1/10000) = SE / sqrt(1/7370 + 1/7270)

SE = 0.0041

d = r-exp - r-cont = 565/7270 - 352/7370 = 0.03

m = 0.0041 \* 1.96 = 0.008

CI lower = 0.0300 - 0.0080 = 0.022

CI upper = 0.0300 + 0.0080 = 0.038

Given that the confidence interval boundaries are not including 0, the effect size test is significant.

**Result of sign test**

Given the daily data, out of the 7 days every day the experiment group shows better results than the control group. *H0: there is no significant difference between the result of control and experiment group.*

*H1: there is significant difference.*

Thus the questions can be turned into calculating the probability of rejecting H0 given significance level of 0.05.

The two tailed P value of binomial test is 0.0156 < 0.05. Thus the sign test is statistically significant. Reject H0 and in favor of H1.

Since there is significant difference between control and experiment group, the decision is to launch the new button.

**Simpson's paradox**

<https://en.wikipedia.org/wiki/Simpson's_paradox>

The effect that a trend appears in different groups but disappears when the groups are combined together.

**Example scenario**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment** | **N-cont** | **X-cont** | **CTR-cont** | **N-exp** | **X-exp** | **CTR-exp** |
| **Experienced users** | 1000 | 30 | 0.03 | 5000 | 200 | 0.04 |
| **New users** | 5000 | 500 | 0.1 | 1000 | 110 | 0.11 |
| **Total users** | 6000 | 530 | 0.088333333 | 6000 | 310 | 0.051666667 |

In this example, for both experienced and new users separately the click-through-rate is higher in the experiment group than in the control group. However, when combining the experienced and new users together, the click-through-rate is higher in the control group than in the experiment group.

**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

Statistical hypothesis testing is based on rejecting the null hypothesis if the likelihood of the observed data under the null hypotheses is low. If multiple comparisons are done or multiple hypotheses are tested, the chance of a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a Type I error) increases.The Bonferroni correction compensates for that increase by testing each individual hypothesis at a significance level of alpha-overall / n, where alpha-overall is the desired overall alpha level and n is the number of hypotheses. For example, if a trial is testing 20 hypotheses with a desired alpha of 0.05, then the Bonferroni correction would test each individual hypothesis at 0.0025.

alpha-individual = alpha-overall / n

* Simple
* No assumptions
* Conservative, guaranteed to give alpha-overall at least as small as specified

**Additional resources**

<https://en.wikipedia.org/wiki/Bonferroni_correction>

<https://en.wikipedia.org/wiki/Multiple_comparisons_problem>

<https://en.wikipedia.org/wiki/Boole%27s_inequality>

<https://en.wikipedia.org/wiki/False_discovery_rate>

<https://en.wikipedia.org/wiki/Closed_testing_procedure>

1. When to use pooled standard error?

2. How to calculate pooled standard error?