# Lesson 5 Analyzing Results

## Outline

1. Sanity Checks
2. Single Metric
3. Multiple Metrics
4. Gotchas

## Sanity Checks

Check invariant metrics for a valid experiment. There are two types of checks:

* Population sizing metrics, based on unit of diversion. This ensures that control and treatment populations are actually comparable.
* Actual invariants, which are not supposed to change during our experiment.

## Choosing invariant metrics

|  |  |  |
| --- | --- | --- |
| **Candidates** | Change order of courses in course list  (unit of diversion: user id) | Change infrastructure to reduce load time  (unit of diversion: event) |
| # signed-in users | Good for population sizing.  Users are randomly assigned to each group | Good for population sizing.  Larger than unit of diversion, one user can have multiple random events. So should be similar between groups. |
| # cookies | Good for population sizing.  Not directly randomized but should be split evenly. Unless users in the experiment tend to clear their cookies more often or visit fewer pages. | Good for population sizing.  Larger than unit of diversion, one cookie can have multiple random events. So should be similar between groups. |
| # events | Good for population sizing.  Not directly randomized but should be split evenly. Unless users in the experiment tend to clear their cookies more often or visit fewer pages. | Good for population sizing.  Events are randomized to each group. |
| CTR on “Start Now”  (between homepage to course list) | Good for actual invariant.  CTR happens before course list, therefore it shouldn’t be affected by this change. | Good for actual invariant.  CTR happens before viewing any videos. There could be a learning effect, but we won’t catch learning effect by event anyway. |
| Time to complete | Not good.  Could be affected. If ordering the courses differently does cause users to enroll in different courses, then that could change how long it takes users to complete the courses that they enrolled in. | Not good.  We can’t track this metric by event, since same user can be assigned to both the experiment and the control group multiple times during the experiment. Besides load time can change how long it takes to complete a class. |

**Another example**: Change location of sign-in button to appear on every page. (unit of diversion: cookie)

The sign-in button currently appears on the course list page, and if a user who isn’t signed in tries to enroll in a course, they are prompted to sign in. But in the experiment group, the sign-in button is added to every page, including the home page.

Which metrics would make good invariants?

1. Events?

A good population sizing metric. # cookies, and # users are good as well. # events should be split evenly by groups, but if it turned out different, it would be good to catch.

1. CTR on “Start Now”?

“Start Now” is on homepage. Adding sign-in button to homepage could affect this. Maybe fewer people would click the “Start Now” button if they instead sign-in and go directly to a course they had already started.

1. Probability of enrolling?

Not good. Users often enroll after signing in.

1. Sign-in rate?

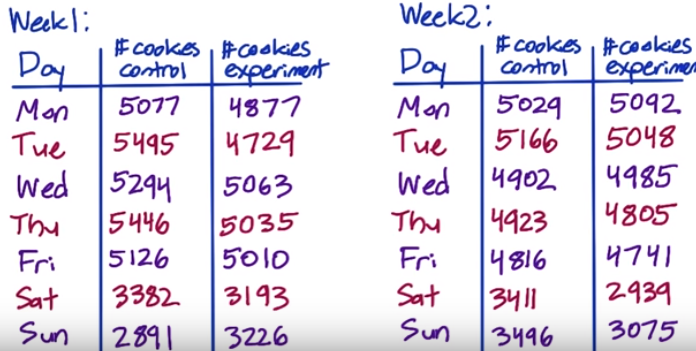
This is actually we are trying to change and measure.

1. Video load time?

A good actual invariant metric since there is no backend change.

## How to check invariants?

If we run experiment for 2 weeks and the unit of diversion is cookie.



Total cookies in control is 64,454 and total cookies in experiment is 61,818. How can we figure out whether this difference is within expectation? Given each cookie is randomly assigned to the control or experiment group with probability of 0.5.

The hypothesis is the proportion of success is 0.5, .

1. Compute standard error of the proportion of success, which is 0.5.
2. Multiply by z-score to get margin of error for 95% CI
3. Compute confidence interval around 0.5.

Which means 95% of the time, the observed fraction of cookies should fall within this range.

1. Check whether observed fraction is within interval.

Which is outside of 95% CI, meaning the observed value is significantly different from 0.5 success rate.

Therefore, there is something wrong in the setup. To get an idea of what could be going wrong, it’s a good idea to look at the day by day data again.

## Side Note: Binomial distribution

There are two types of binomial distribution:

* Distribution of number of successes,
* Distribution of the proportion of successes,

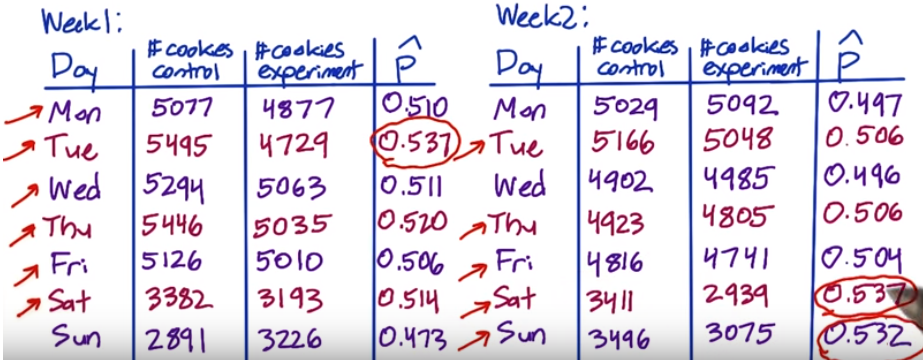
Sampling variance is:

Bur since we don’t know what the actual probability of success , we use for estimation (CLT). Therefore,

The estimated sampling variance is:

The estimated standard error is:

## Checking invariants and finding causes



In 11 out of 14 days, there were more cookies in the control group than in the experiment group, which is quietly high. And if we also compute the fraction of the cookies that were in the control group for each day, there were a few days with 0.53 or higher. No day stands out as an obvious problem. This points to an overall problem rather than a problem on a specific day.

**What to do?**

* Talk to the engineers and figure out if something was wrong with the experiment setup.
* Try slicing to see if one particular slice is weird. For example, by country, by language or by platform to see if one particular slice looks like it’s causing the problem.
* Check the age of cookies – does one group have more new cookies?

## What happens next when sanity check fails?

Do not proceed to, go straight to analyzing why sanity checks fail. A few things to check:

* Something might have gone wrong technically. We should work with engineers and see if there is something wrong with the experiment infrastructure, such as setup or diversion.
* Retrospective analysis. Try and recreate experiment diversion from the data captured, and understand whether there is something endemic to what we’re trying to do that may be causing the situation.
* Compare to pre-period A/A tests. If changes in invariant metrics still exists, that could point to an infrastructure problem. If not in the pre-period, that would mean we have problems with data capture during experiment.

What’s the most common reasons for failing the sanity check?

* Data capture, especially when we want to capture a new experience that the user is undergoing.
* Setups. Maybe we only have an English filter for experiment but not for control.

## Single Metric