# 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 it 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 it 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?

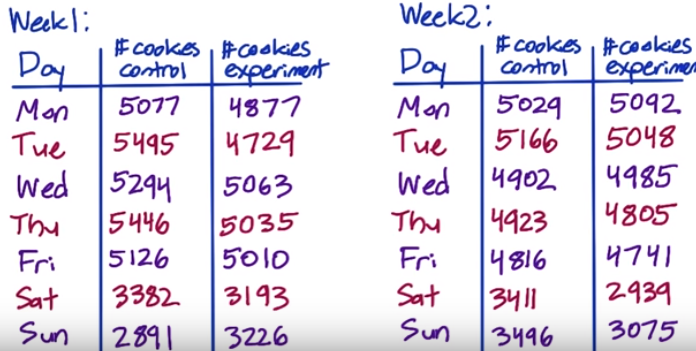
This is actually we are trying to change and measure.

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,

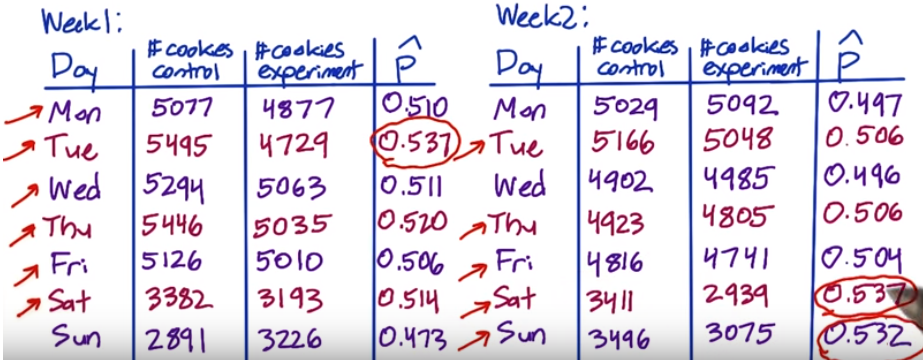
Sampling variance is:

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

The goal of the analysis is to make a business decision about whether our experiment has favorably impacted our metrics. Analytically, that means we want to decide if we’ve observed a statistically significant result of our experiment. Typically, we also want to measure the magnitude and the direction of the change. The once we have all that information, we can make a decision about whether we want to recommend our business to launch this experiment or not.

## What’s next if the results aren’t significant?

That’s a good time to take a much deeper analysis at our results, especially if we were expecting a really noticeable difference. For example, we might want to break it down to different platforms or different days of the week, which can not only help us to find bugs in our experiment setup, but it might give you a new hypothesis about how people are reacting to the experiment.

Also, if this is our first go around, we may also want to try cross checking our results with other methods, such as comparing results from non-parametric sign tests and parametric tests.

## What not to do if the results aren’t significant?

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 were expecting! 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.

## Analysis with a single metric

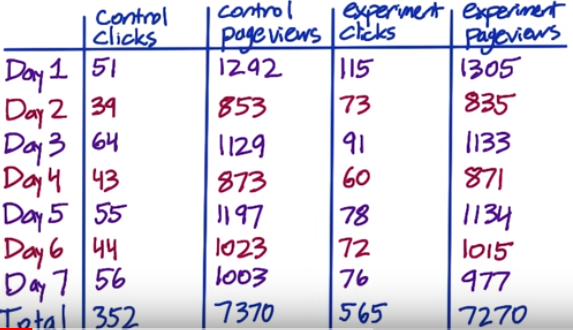
Experiment: change color and placement of “Start Now” button

Metric: click-through-rate

Unit of diversion: cookie

* The practical significance boundary (the smallest effect that will be detected (1-β)% of the time.)
* The significance level is (Percent of the time a difference will be detected, assuming one does NOT exist)
* The statistical power is (Percent of the time the minimum effect size will be detected, assuming it exists)

Here is the result:



After we determine that sanity checks pass, which means the number of page views is comparable between the two groups, let’s start analyzing.

Click-through-rate is more like to be Poisson distribution, while the click-through-probability is binomially distributed. Therefore, we will estimate the variance of CTR empirically. In fact, it was already calculated during the experiment design:

Since SE is proportional to the inverse of square root of sample size,

Or

Therefore, the standard error for our experiment sample will satisfy:

Which gives us

Now we have,

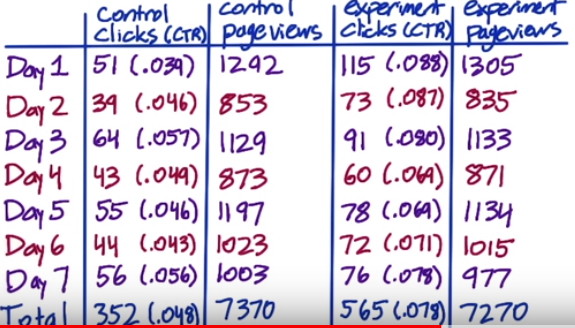
The difference between two groups is:

The margin of error for a 95% confidence level is,

The confidence interval is,

Which doesn’t include our practical significance boundary 0.01, which means at a 95% confidence level, we can be confident that the true change is large enough to be worth launching.

However, just to double check, let’s look at the results of the sign test and bring up the day by day data with CTR for each group.



# of days: 7

# of days with positive change: 7

We want to know the probability of this happens due to random chance. If there is no difference between control and treatment groups, the chance of positive change on each day will be 50%. So, the question is, if we flip a fair coin seven times, what is the chance it comes up heads seven times? Using a two-tailed test,

Under the null hypothesis, this result is unlikely to come about by chance.

Given the above analysis, we will still recommend launching this change.

## What to do when the hypothesis on the effect size disagrees with the sign test?

For example, the hypothesis on the effect size is significant while the sign test is not.

Why this happens?

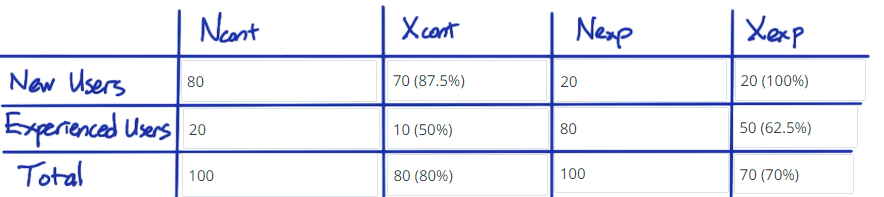
The sign test has lower power, which is typical for non-parametric test. Why? Binomial distribution vs normal distribution assumption in parametric test. That’s the price for making no assumptions, less chance to detect the difference assuming it exits. But this isn’t necessarily a red flag, but it’s worth digging deeper and seeing if we can figure out what’s going on.

* Break down into weekday and weekend, and construct CI for each scenario and see whether they’re significant for both tests now.
* Compare to practical significance level, talk to manager and see if it’s worthy launching.

## Simpson paradox

There’s a bunch of different subgroups in our data, such as user populations (new users vs. experienced users). And within each subgroup, we can observe a certain directional trend or relationship, but the same trend is not observed (or sometimes reverse) when we look at the combined dataset. This is commonly observed when analyzing/comparing proportions or averages (e.g., conversion rate and average booking value).

Here is an example:



In both new users and experienced users subgroups, experiment group has higher CLT than control group, but not on the total data. This happens because new users also have a higher overall CLR than the experienced users, which explains why control group (which has more new users than experiment group) has a higher CLR overall.

But why are there more pageviews from new users in the control group? By randomization design, shouldn’t they be the same?

As a matter of fact, they should. During our sanity check, it’s a good idea to make sure the sample size is the same in both control and experiment groups. Checking that breakdown across different slices could also be a good sanity check.

Reasons for Simpson paradox to happen:

* Something wrong with the set-up
* Or the change we made affects new users and experienced users differently. For example, maybe the change makes experienced users refresh the page more often than new users, which causing more pageviews (samples) for experienced users in the experiment group.

Even though it’s attempting to say this is a successful experiment since it was an improvement for both users and experienced users, we really need to dig deeper and figure out why there are more samples from experienced users in the experiment group. No matter what the cause is, we won’t be able to make a valid conclusion until we understand what’s going on.

## Multiple metrics

When we run evaluations of multiple metrics at the same time, the more things we test, the more likely we are going to see significant differences just by chance. For example, if we’re testing 20 metrics, and we have a 95% confidence level, we would expect to see one case at least that time when we got the significant result but it’s only concurring by chance.

This is a problem, but we’re not sunk, because it shouldn’t be repeatable. That is if we do the same experiment on another day or we divide data by slices or we do bootstrap analysis, we wouldn’t see the same metric showing up as significant differences every time. It should occur randomly.

We can use multiple comparisons to adjust the significant level, so it accounts for how many metrics or how many different tests we’re doing. We can do automatic detection of differences to alert us that some metric is unusual higher or lower, so we can perform multiple comparison instead of single comparison.

## Tracking multiple metrics example

Experiment: prompt students to contact coach more frequently

Metris:

* Probability that student signs up for coaching
* How early students sign up for coaching, e.g. the average amount of progress a student makes before enrolling for coaching.
* Average price paid per student

If Audacity tracks all three metrics and does three separate significance tests (), what is the probability of at least one metric will show a significant difference if there is no true difference? (at least 1 false positive)

There is an independent assumption for these 3 metrics, which is not true. 3 metrics tend to correlate to each other and move together, therefore 0.143 is an overestimate of the probability of a false positive. But assuming independence is an easy way to get a conservative estimate.

For n metrics with , the overall probability of at least 1 false positive is (assuming independence):

For example,

* 10 metrics with 95% confidence, overall is 0.401. Almost half chance.
* 10 metrics with 99% confidence, overall is 0.096. Still higher than 0.05.

## Bonferroni correction

Problem: the probability of any false positive increases as we increase the number of metrics.

Solution: use higher confidence level for each metric.

**Method 1**: assuming independence,

Then solve for .

**Method 2** (more practical): Bonferroni correction

* Simple to calculate
* No assumptions
* Conservative: guaranteed to give at least as small as specified.

Sometimes Bonferroni correction is too conservative, there are some less conservative strategies:

* Closed testing procedure
* Boole-Bonferroni bound
* Holm-Bonferroni method

Intuitions behind different strategies:

* Control probability that any metric shows a false positive

, also called the familywise error rate (FWER)

* Control false discovery rate (FDR)

This only makes sense if we have a lot of metrics, let’s say 200 metrics, if we cap FDR at 0.05, this means we’re okay with 5 false positives and 95 true positives in **every experiment**.

## Drawing conclusions

Once we’ve figured out which metrics have significant changes, what comes next?

Now we have to decide what our results do and don’t tell us. If we have statistically significant results, then that means that we’re unlikely to have zero impact on the user experience. But now the questions come down to:

* Do we understand the change?

Some metrics show significance while some other metrics don’t. Having intuition and experience with lots of other experiments can really help here. For example, maybe we know that for small changes, a change in one metric but no change all the other metrics is perfectly fine. But if we saw the same results for a big change, that would probably indicate that there’s something going wrong.

* Do we want to launch the change?
* Do we have both statistically significant and practical significant results in order to justify the change?
* Do we understand what that change has actually done with regards to user experience?
* Is it worth it?

## Changes over time

It’s a good practice to ramp up our experiment over time, for example, maybe we start with 1% of our traffic as being diverted to the experiment group, but then we gradually increase that until our feature is fully launched. Also we should remove all the filters and test the change during our ramp-up on all users.

But there is one gotcha: the effect may actually flatten out as we ramp up the change. Here are a few reasons:

* Seasonality effects. Students, for example, when students go on summer vacation, the behavior across broad swaths of the Internet changes. Or holidays, shopping behaviors can change.

One way to capture these seasonal or event-driven impacts is holdback. We launch our change to everyone except for a small holdback. This set of users don’t get the change and we continue comparing their behavior to the control. In that case, we will see a reverse of the impact that we saw in our experiment. We can also track that over time until we’re confident that our results are actually repeatable.

* Novelty effect or change aversion. User behaviors can change as we launch the new feature. Cohort analysis can be helpful.
* If advertisers have budgets and we don’t control for the budgets properly, the effect can change as we ramp up.

If we concern about learning effect, we can use pre- and post-periods, combined with cohort analysis to study how users are adapting to the change over time.