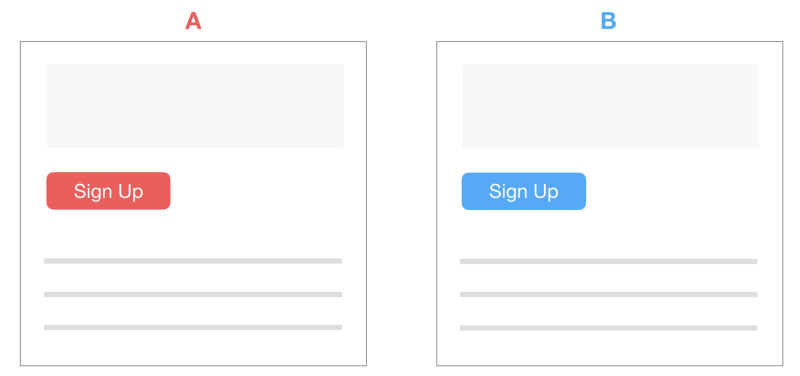
**An Introduction to Bayesian A/B Testing in Exploratory**

What is A/B Testing?

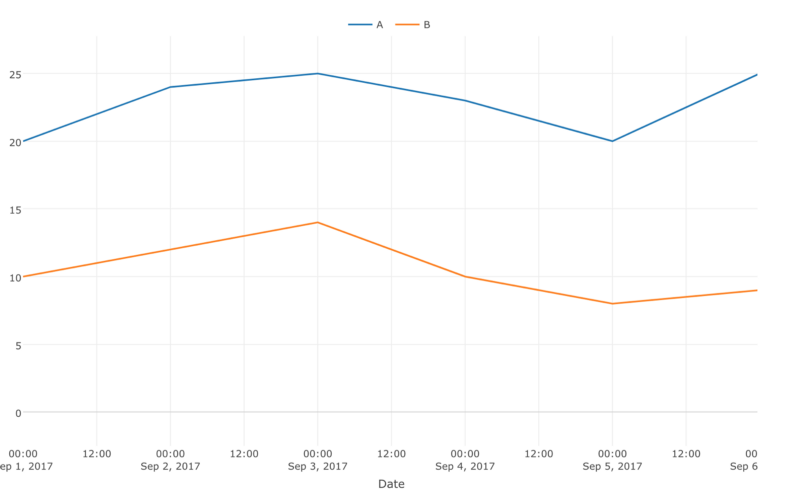
<https://blog.exploratory.io/an-introduction-to-bayesian-a-b-testing-in-exploratory-cb5a7ad80963>

I’m sure many of you have heard about A/B Testing, and maybe some of you have done it before or even doing it right now. Basically, you create two groups — A and B — and measure the performance of each group and compare them to see which one is better.

It is a very well known technique especially among web / mobile developers or product people. For example, you deploy two versions of your web site landing pages and measure which pages are helping your users to sign up, subscribe your service, etc.

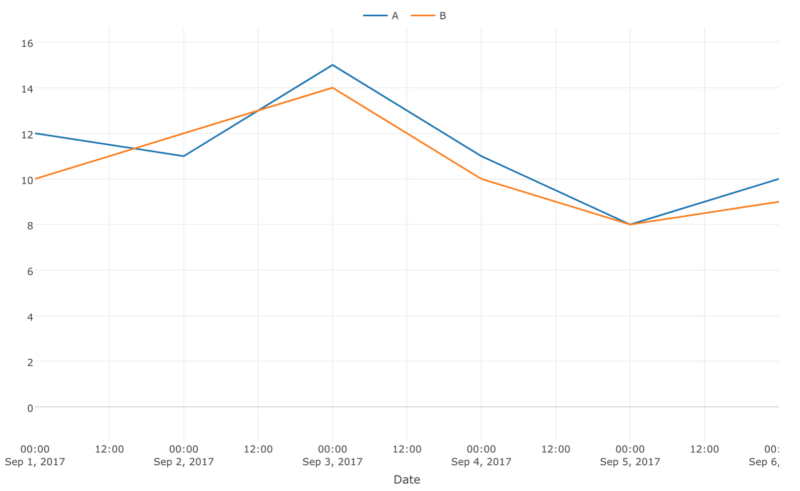


Now, let’s say your test result came back after a week or so and it looks like below.



This is easy. A (blue color) is consistently performing much better than B (orange color) so you can deploy A version of the page to the production. Done.

Except, it is not that simple in most of the times. The difference between A and B can be very subtle and it can look something like below.



Just by looking at this, you might think that A seems to be better than B. But you might not be confident enough because B was better than A in one day and the difference was kind of small even for the days where A was better than B.

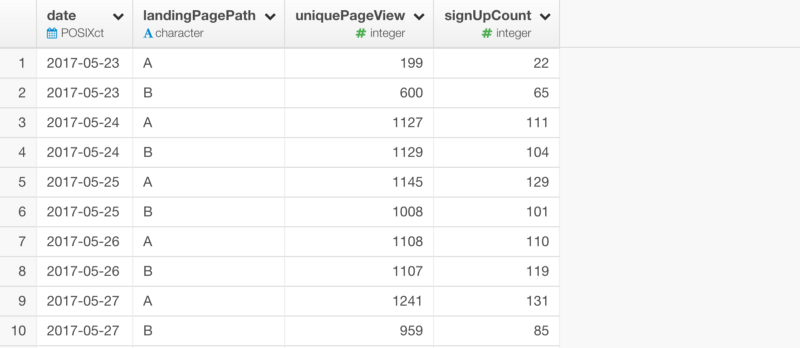
Now, would you be comfortable making a decision to go with A?Are you sure to say that B won’t be performing better than A tomorrow? Because deciding to go with A means that you will most likely invest an additional time and money in the development, the design, the deployment, and all the other related works. So you would really want to be certain that A was indeed better than B.

This is where the power of Statistics comes in. There are several ways to evaluate the result with some techniques often used in Statistics. Traditionally, the most popular one is to use something called [Chi-Squared Test](https://en.wikipedia.org/wiki/Chi-squared_test).

Let’s quickly take a look at how this can be done.

**Chi-Squared Test**

Let’s say we are testing two versions of our landing page and monitoring how much ‘sign ups’ each of the pages is bringing in every day.



Now you can summarize the values by the landing page.



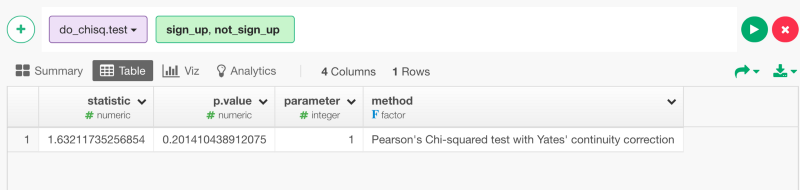
Note that this test doesn’t really care about the trend by the date. It’s about how much of the data we have. In reality, we need to first estimate how much data we need to collect before we run this test, but for a sake of simplifying this, we are assuming that we have collected just enough data already.

Also, note that I have calculated ‘not\_sign\_up’ column as

total - sign\_up

because Chi-Square Test wants to see the proportion of the data.

Now we can run the test and get a result like below.



The most important thing we want to look at here is ‘p.value’ column. This shows the rate that this difference between A and B can happen by a random chance. In this case, the number is 0.2014, which means that this difference can happen by chance at a rate of 20%. And this means, in statistic world where this value needs to be less than 5% to state a statistical significance, we can’t conclude that the difference between A and B is statistically significant. Yes, the statistics can’t support you make any decisions here!

Now, Maybe you might ask as a business person, what if we continue the testing and collect more data, then we’ll evaluate the result again? We might see the P-value going down to less than 5% at some point? The answer is No. Unfortunately, the test is already done. The reason we have run the test is that we have collected enough data for the test to be valid. That means, more data is not going to help the test. The only thing we can do is to re-design the test by starting from scratch. Ouch!

Along with this problem, there are a few challenges with using Chi-Square Test for A/B Testing especially for the modern organizations who need to make decisions quickly and iteratively.

We need to know how much of the data we need to collect for the test before starting the test.

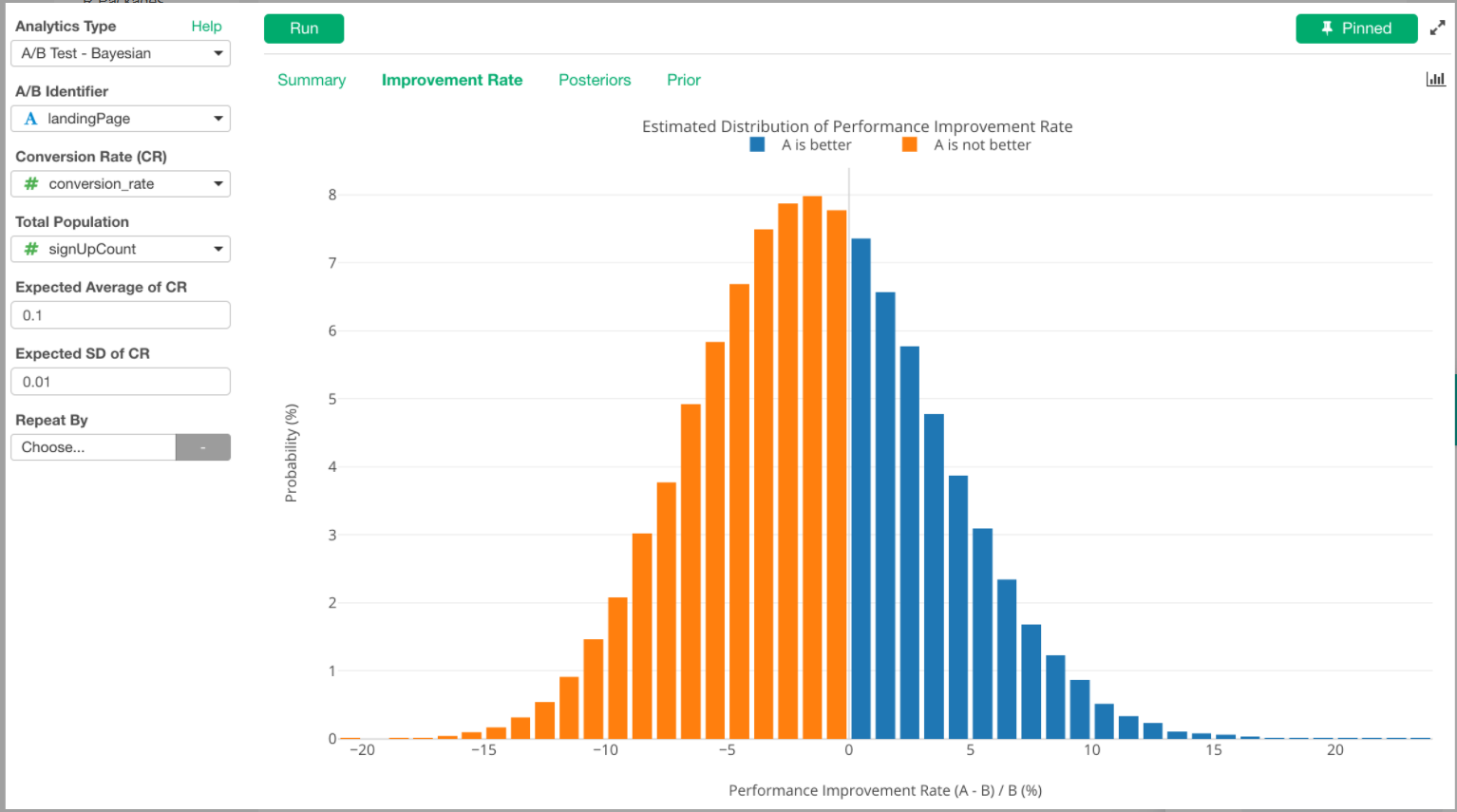
We can’t test the result in real-time until we collect a full of the planned data size.

The test result is not intuitively understandable especially for those without a statistical background.

1. The test result is black and white, either it is statistically significant or not. Therefore, it’s hard to figure out what to do especially when it is not statistically significant.

From Chat:

Here's how you can assign the columns for A/B under Analytics view.



And by using Pivot table under Viz, you can calculate the Average / SD (Standard Deviation) quickly.

