Published in Towards Data Science

Kaushik Sureshkumar Follow
Jan 26, 2021 · 8 min read · ✦ · Listen

Bayesian AB Testing — Part III — Test Duration

Photo by Aron Visuals on Unsplash

Series Structure

This post is the 3rd part of a series of blog posts on applying Bayesian AB Testing methods to real life product scenarios. It uses some of the concepts discussed in the 1st part of the series.

1. Modelling and analysis of conversion based test metrics (rate metrics)
2. Modelling and analysis of revenue based test metrics (continuous metrics)
3. Calculating test duration
4. Choosing an appropriate prior
5. Running tests with multiple variants

Experiment Context

Following on from the example used in a previous post, let's assume we've recently changed the messaging on an upsell screen and want to AB test it before releasing to our wider user base. We hypothesise that the changes we've made will result in a significantly better conversion rate.

We go ahead and model our conversion rate as a Bernoulli random variable with conversion probability λ, which in turn we model with a prior distribution of $Beta(7,15)$. We then choose our expected loss threshold of $\epsilon = 0.0015$. We're now ready to run our test, but how long do we run it for?

Before we dive into the analysis to answer this question, let's first consider what test duration actually means. A common method for calculating test duration is given by the following formula:

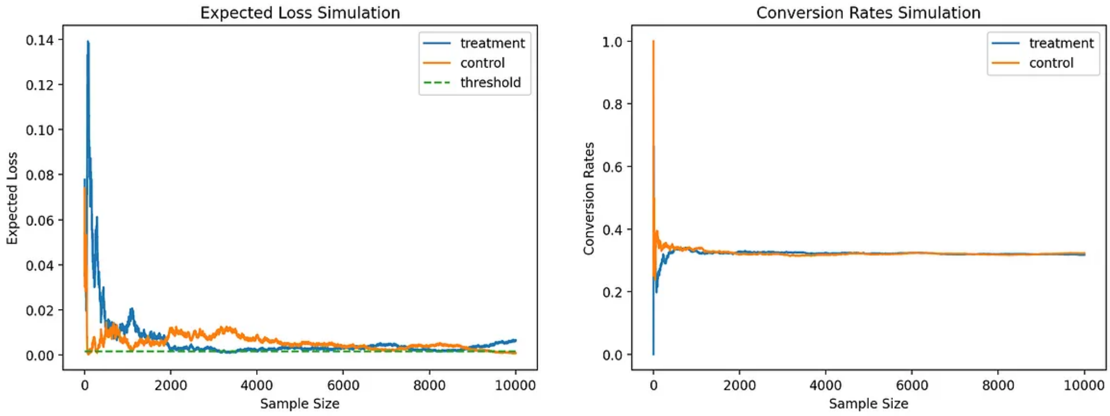
$$test\ duration\ (weeks) = \frac{sample\ size\ required\ per\ variant\ *\ no\ of\ variants}{expected\ no\ of\ weekly\ active\ users\ *\ proportion\ of\ users\ to\ be\ included\ in\ the\ test}$$

The number of variants and proportion of users to be included in the test are factors that would've been decided during the designing of the test. The number of users we expect to have in a week is a simple calculation based off data from previous weeks. So essentially, the problem boils down to choosing a sensible required sample size for each variant in the experiment.

It is often suggested that bayesian product experiments should be run until the expected loss of one of the variants drops below our threshold ϵ, at which point we declare that variant the winner. This way we don't need to worry about calculating the required sample size. However following this approach could result in us wrongly choosing a variant to roll out due a concept known as peeking [1]. To explore this further, let us consider the following example.

Peeking

Let us consider the case where λ_c and λ_t are pretty much identical. We simulate running an AB test and expect the results to be inconclusive. However we see the following result.



Peeking Example (Image by Author)

We see from the conversion rates on the right that λ_c and λ_t converge within the first 1000 samples. The more interesting observation, however, is to do with the graph on the left. We see that within the first few samples, the control expected loss falls below the threshold. If we were to stop the test here, we'd conclude that control was significantly better than treatment, which would be the wrong conclusion. We also see that if we don't stop the test here, depending on when we stop the test, all three results (control win, treatment win and inconclusive test) are possible. So how do we decide when to stop the test?

Running simulations of our experiment will help us avoid falling into the trap of peeking and will ultimately help us choose a sensible required sample size per variant. Before we look at how this works let us consider the following caveat.

Caveat

The main caveat for the proposed method below is that it doesn't take into account any experimental design arguments for choosing a test duration. In particular it doesn't take into account any seasonal or time based variation in conversions. In the real world, it is pretty likely that the conversion which you use as your test metric will vary depending on the day of the week and even time of the day. As such, it's worth running the test for at least a couple of weeks to avoid drawing the wrong conclusions due to seasonal variation. I'd suggest using the method outlined in this post to calculate how much longer than 2 weeks to run the test for.

Sample Size Calculation

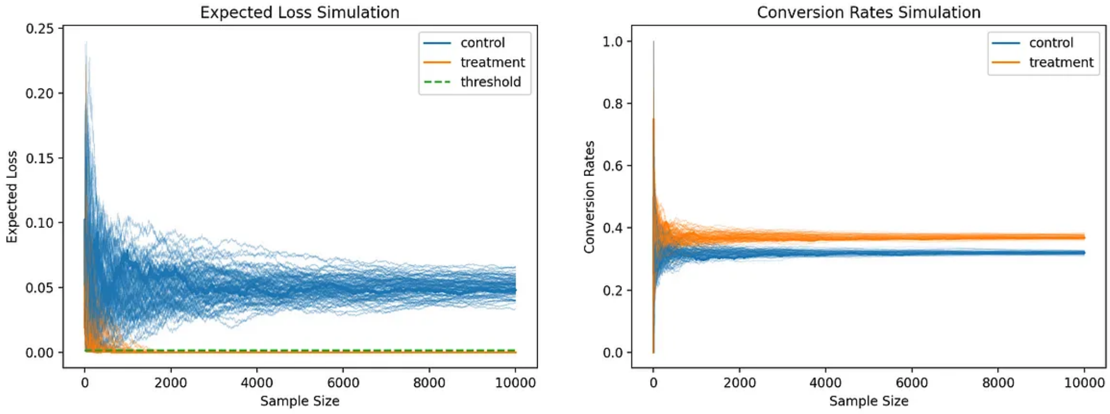
There are three main factors which affect the required sample size:

- Minimum detectable effect δ— The minimum change in λ we need to roll out treatment
- Expected loss threshold ϵ — The maximum loss in conversion rate we're willing to accept in the case where we wrongly declare a winner
- Scale of conversion probability λ

We will be diving deeper into how each of these factors affects sample size later on in the post, but let us first establish a process for calculating a sensible sample size.

Let us first assign some values to the factors above. We run 100 simulations of the experiment with an average prior conversion rate of 32%, expected loss threshold ϵ of 0.0015 and a relative minimum detectable effect of 15% (so we're looking for $\lambda\_t \geq 0.15 * \lambda\_c$).

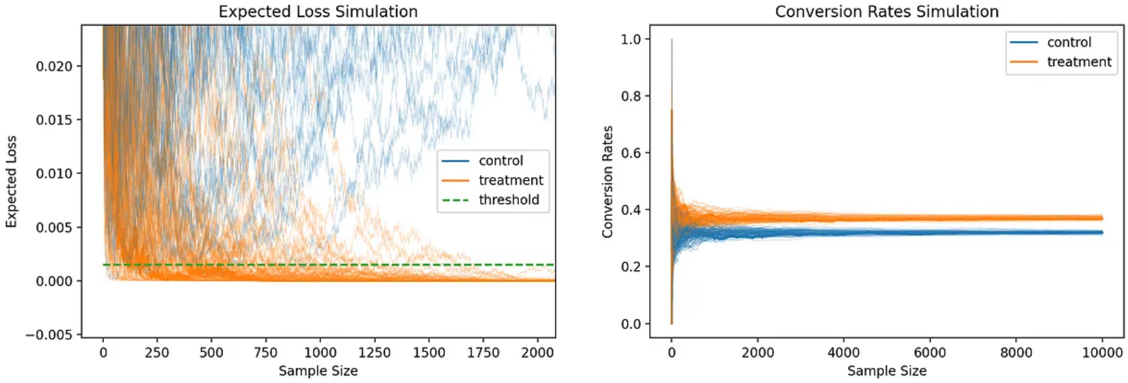
Let's now explore the simulated data.



Experiment Simulations (Image by Author)

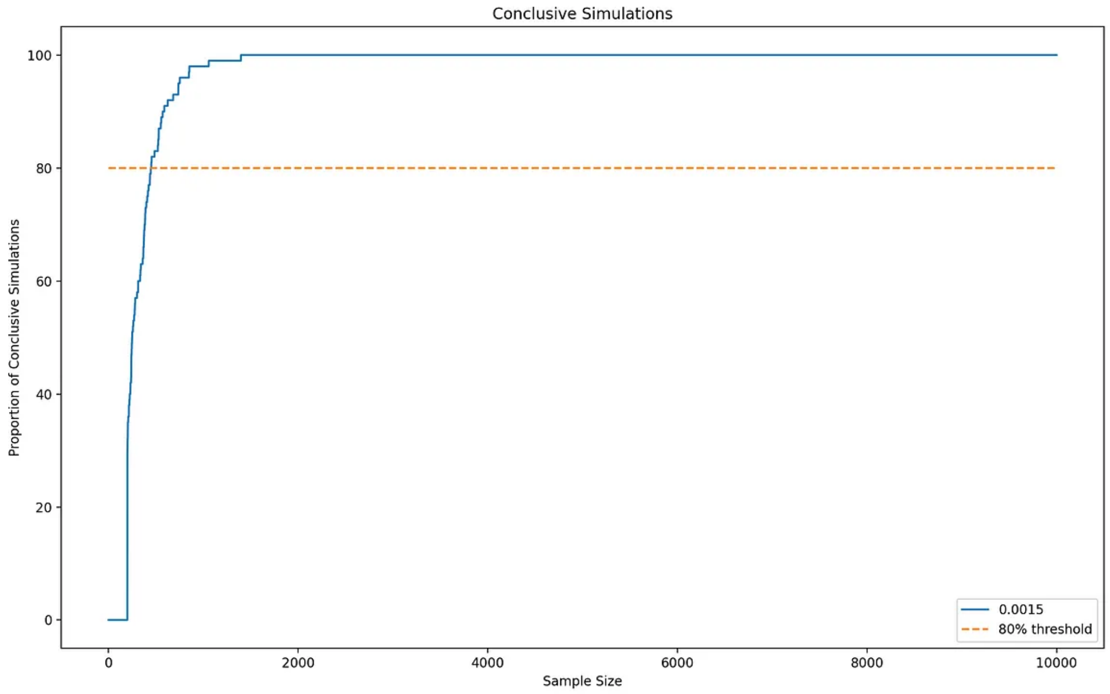
By inspection, we can see that in the case where the minimum detectable effect is achieved by treatment, most experiments are concluded by the time we get to 2000 samples per variant. However this is still 4000 users we need to expose to the test which is a lot and may take a while. So let's see if we can reduce this while still being pretty sure that the test would be conclusive.

In order to stop the test we need to set a minimum number of samples so that we don't fall into the trap of peeking. Let's zoom into the expected loss graph above to choose this minimum number of samples.



Experiment Simulations — Zoomed In on Low Sample Size (Image by Author)

Once again, by close inspection, we can see that most false positives — cases where the expected loss of choosing control drops below the threshold — can be eradicated if we set the minimum sample size to 200. We now go ahead and plot how many samples are required to declare a winner in each simulation [2], given that at least 200 samples are observed per variant.



Conclusive Simulations (Image by Author)

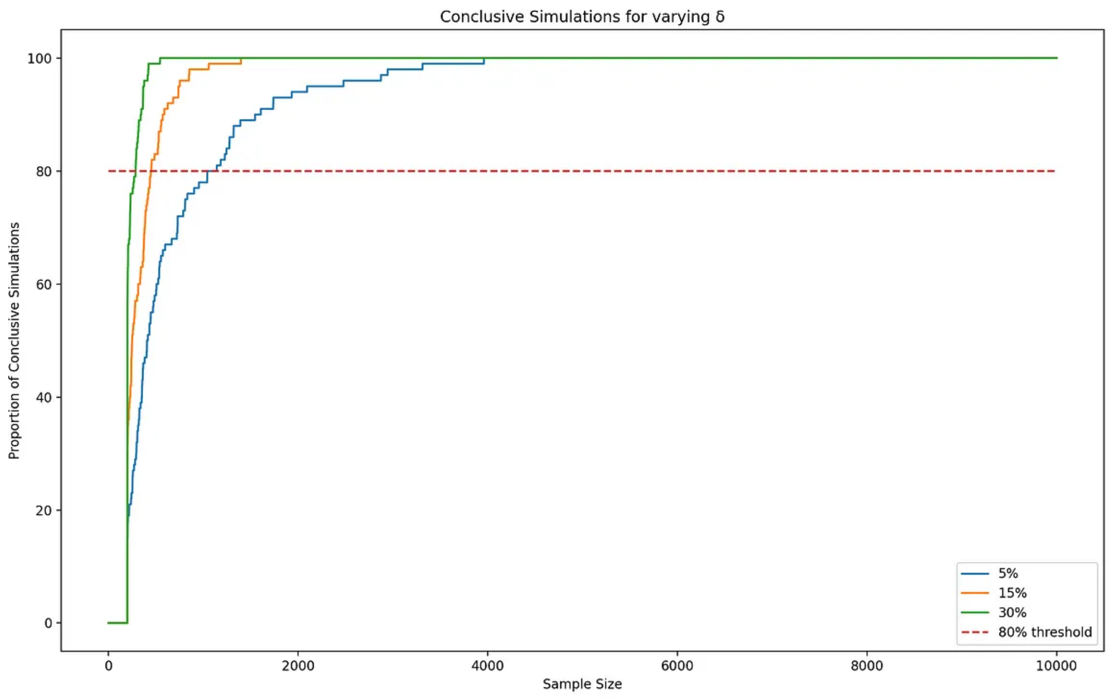
```
For expected loss threshold of 0.0015, minimum detectable effect of 15% and prior avg conversion rate of 32%, 80% of
simulations needed 447 samples per variant to be conclusive
```

We see that we only need 450 samples per variant for 80% of the tests to be conclusive. So if we decided to run the test until we had 900 samples then, given that the minimum detectable effect is achieved, we have an 80% chance that the test would be conclusive. Thus we see that we'd be pretty likely to find a conclusive result with 3100 fewer samples than we originally thought.

Applying the same logic, let's now look at how the required sample size is affected by the difference factors we discussed earlier.

Minimum Detectable Effect

The minimum detectable effect δ is the smallest relative percentage change in λ that we want to detect. It's the smallest change that will make the treatment worth rolling out to our user base. Let's have a look at how this δ affects our sample size calculations. We're going to use a prior average conversion rate of 32% and an expected loss threshold of 0.0015 while choosing δ from {0.05,0.15,0.3}.



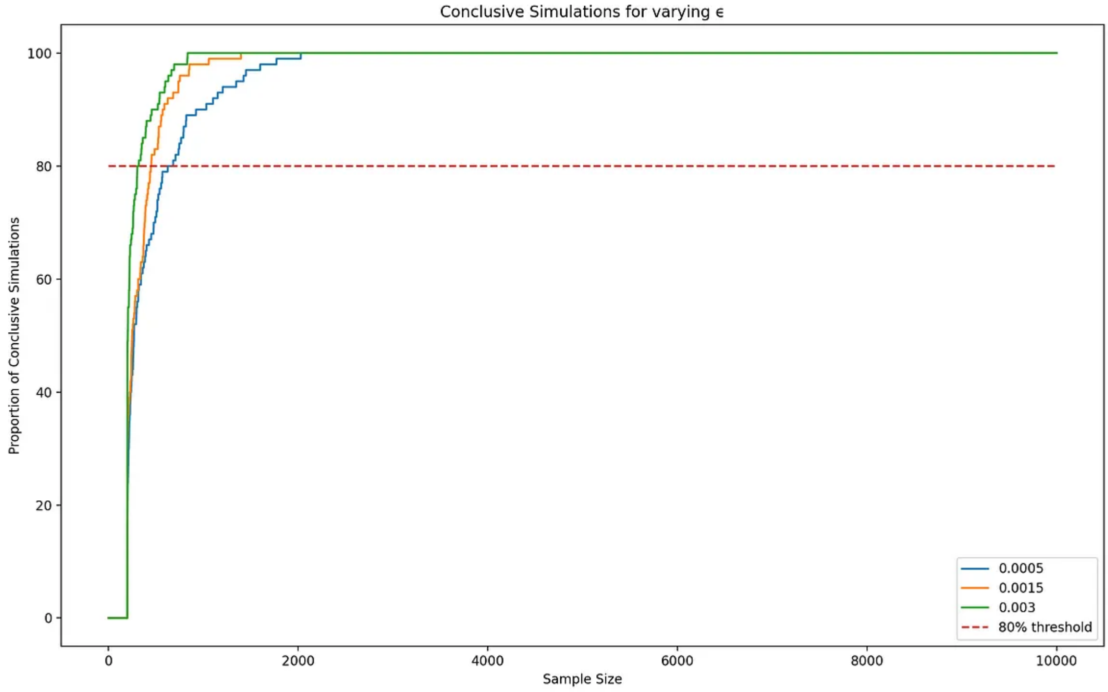
Conclusive Simulations for Varying δ (Image by Author)

```
For minimum detectable effect of 5%, 80% of simulations needed 1043 samples per variant to be conclusive
For minimum detectable effect of 15%, 80% of simulations needed 447 samples per variant to be conclusive
For minimum detectable effect of 30%, 80% of simulations needed 286 samples per variant to be conclusive
```

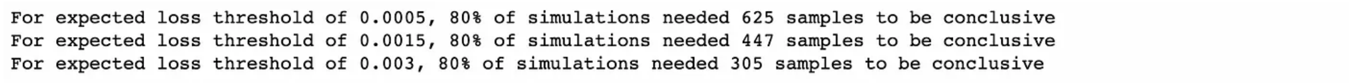
We see that the higher the minimum detectable effect, the lower the required sample size per variant for 80% of tests to be conclusive and vice versa. Choosing a δ of 5% would require us to get about 2100 samples in total, whereas choosing a δ of 30% would require us to only get about 600 samples in total. Intuitively, this makes sense, since the larger the change in conversion probability the fewer the samples we'd need to be sure of it.

Expected Loss Threshold

The expected loss threshold ϵ is the maximum expected loss we're willing to accept in the case where we mistakenly choose a variant. It is the maximum expected drop in conversion rate we'd be happy with in this scenario. Let's have a look at how this ϵ affects our sample size calculations. We're going to use a prior average conversion rate of 32% and minimum detectable effect of 15% while choosing ϵ from {0.0005,0.0015,0.003}.



Conclusive Simulations for Varying ϵ (Image by Author)

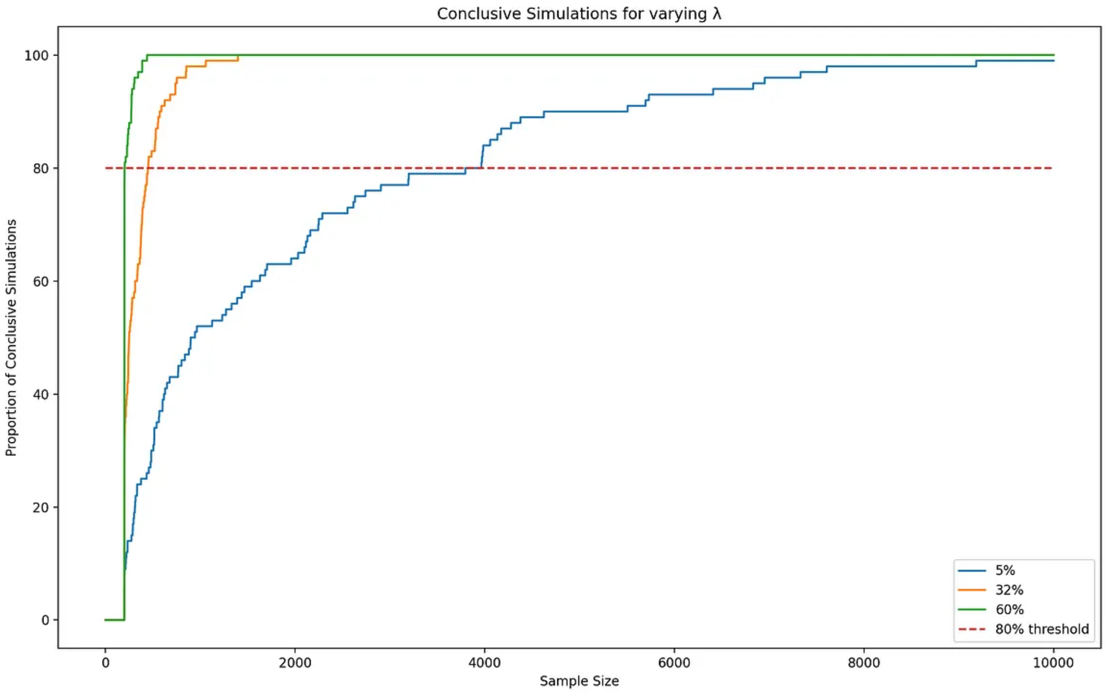


```
For expected loss threshold of 0.0005, 80% of simulations needed 625 samples to be conclusive
For expected loss threshold of 0.0015, 80% of simulations needed 447 samples to be conclusive
For expected loss threshold of 0.003, 80% of simulations needed 305 samples to be conclusive
```

We see that the higher the expected loss threshold, the lower the required sample size per variant for 80% of tests to be conclusive and vice versa. Choosing an ϵ of 0.0005 would require us to get about 1350 samples in total, whereas choosing an ϵ of 0.003 would require us to only get about 650 samples in total. The changes in required sample size are smaller than the minimum detectable effect case because the scale of change in ϵ is different to the scale of change in δ. However, the concept is similar, the lower the expected loss threshold the more certain we'd like to be of the result of the experiment so the larger the required sample size.

Scale of the Conversion Probability

Let's now have a look at how the scale of the conversion probability used in the test impacts our sample size calculations. We're going to use a δ of 15% and choose a relative ϵ of 0.005. We've used a relative ϵ in this scenario so the analysis is more fair — this is the same as using $\epsilon = 0.0015$ in the previous scenarios where the prior average conversion rate stayed constant as 32%. In this scenario we will be considering prior average conversion rates from {0.05,0.32,0.6}.



Conclusive Simulations for Varying λ (Image by Author)

```
For prior avg conversion rate of 5%, 80% of simulations needed 3796 samples to be conclusive
For prior avg conversion rate of 32%, 80% of simulations needed 447 samples to be conclusive
For prior avg conversion rate of 60%, 80% of simulations needed 202 samples to be conclusive
```

We see an interesting result. Although there isn't a (relatively) big difference in required sample sizes for higher conversion rates, as the conversion rate gets lower the required sample size gets disproportionately large. This is an important concept to bear in mind when calculating required sample sizes. It occurs because the lower the conversion probability λ, the more spread out the posterior distribution of λ will be and thus the more samples we'd need to reduce this spread.

I hope you found this post to be a helpful introduction to estimating a sensible test duration for your bayesian product experiments and understand the factors which affect it. Watch this space for the next part of the series!

References

[1] Is Bayesian AB Testing Immune to Peeking? by David Robinson

[2] Bayesian A/B testing — a practical exploration with simulations by Blake Arnold — I got the idea to look at percentile of conclusive simulations from Blake's post

My code from this post can be found here

Ab Testing Statistics Data Science Product Product Management