Growth Book Bayesian Statistics Engine

Jeremy Dorn[1]

[1]*Growth Book*

July 31, 2021

Abstract

There are many ways to approach A/B test analysis. At Growth Book, we have developed an open source Bayesian statistics engine to provide intuitive results to the three most common questions people have when looking at A/B test results: "Which version is better?", "How much better is it?", and "Do I have enough data to call the test now?" While answering these questions, Growth Book also helps avoid common pitfalls of A/B testing such as sample ratio mismatches and fixed horizon peeking problems.

1 The Problem

A/B test analysis often uses frequentist hypothesis testing methods, such as Chi-Squared or Student T tests. There are two serious issues with these approaches.

First, you must decide the sample size in advance before running an experiment. This is known as a Fixed Horizon. To maintain the validity of the statistics, you cannot stop an experiment early for any reason. Without an early escape hatch, experiments that are clearly winning or losing must continue running for the full duration, increasing risk to your business and reducing your velocity. If you ignore this restriction, you will run into the peeking problem [1], also known as repeated significance testing errors. This happens when you look at results of a frequentist experiment every day (or another interval) and stop when it reaches significance (whether or not the predetermined sample size was reached). Doing this significantly increases your chance of making a Type I error, or thinking an experiment won when it did not.

Second, the results you get from frequentist methods (P-values and confidence intervals) are very difficult for decision makers to interpret correctly. People often incorrectly assume a P-value of 0.05 means that there is a 95% probability the variation is better than the control. The true interpretation is that if the experiment were to be repeated many times, you would only expect to receive a result as extreme as you did 5% of the time. Similarly, people incorrectly assume a 95% confidence interval has a 95% chance of containing the true value. What it actually says is that if the experiment were repeated many times, the fraction of calculated confidence intervals that encompass the true value would tend towards 95%. This is so counter-intuitive and hard to understand that even professional scientists often make these same mistakes in academic papers when dealing with frequentist methods. [2]

2 Bayesian Statistics

With a Bayesian approach, there are no Fixed Horizons. You can look at results frequently and call a test whenever you like and the inferences will still be accurate. [1]

In addition, Bayesian statistics are much easier for people to interpret since it embraces uncertainty [3]. Everything has some probability of being true and you adjust the probabilities as you gather data and learn more about the world. This matches up with how most people think about experiments - “there’s a 95% chance this new button is better and a 5% chance it’s worse.”

2.1 Priors and Posteriors

Bayesian hypothesis testing starts with a Prior distribution that represents what you know about the population before you start your experiment. At Growth Book, we use an Uninformative Prior. This simply means that before an experiment runs, we assume both variations can have any value and have an equal chance of being higher/lower than the other one. If you instead use an Informative Prior, which is based on historical data, it can help with regularization and can shorten experiment times [4]. Growth Book currently doesn’t support Informative Priors, but we plan to add this in the future to make our statistics engine even more powerful.

As the experiment runs and you gather data, the Prior is updated to create a Posterior distribution.

2.1.1 Binomial Metrics

For binomial metrics (simple yes/no conversions), we use a Beta-Binomial Prior with parameters α and β. We use an uninformative prior with both set to 1, which produces a uniform distribution. Given the count of converted users x_A and the total number of users n_A, we can update the Prior to get our Posterior distribution [5]:

$$P_A | X_A \sim Beta(\alpha + x_A , \beta + n_A - x_A) \tag{1}$$

2.1.2 Gaussian Metrics

For count, duration, and revenue metrics, we use a Gaussian (or Normal) Prior with parameters μ_0 , n_0, and σ^2_0. We use a Prior with $\mu_0 = 0, \sigma^2_0 = 1, n_0 = 0$. Given the sample average $\overline{X}_A$ and sample standard deviation s_A, we can update our Prior to get our Posterior distribution by taking a weighted average of the means with weights inverse to their variances: [5]

$$\mu_A | X_A \sim N\left(\left(\frac{n_A}{s^2_A} + \frac{n_0}{\sigma^2_0}\right)^{-1} \left(\frac{n_A}{s^2_A} \cdot \overline{X}_A + \frac{n_0}{\sigma^2_0} \cdot \mu_0\right), \left(\frac{n_A}{s^2_A} + \frac{n_0}{\sigma^2_0}\right)^{-1}\right) \tag{2}$$

You may notice that some of the metric types we are treating as Gaussian are not typically normally distributed (e.g. revenue is usually highly skewed towards low amounts). Luckily, the Central Limit Theorem (CLT) applies since we are actually dealing with the distribution of the sample means and not the raw data itself [6]. Extremely skewed distributions may not follow the CLT fast enough for a given site's traffic levels. In that case, we support setting "caps" on a metric's value. For example, if your typical order value is $5, but every once in a while you get a $10,000 bulk order, you can set the cap to $100 to prevent those extreme outliers from messing up the distribution too much. Capped values plus the CLT ensure almost every continuous distribution can be modeled accurately with a Gaussian model.

3 Common A/B Test Questions

When looking at A/B test results, there are three questions that usually come up:

1. Which version is better?
2. How much better is it?
3. Do I have enough data to call the test now?

These can be answered by calculating inferential metrics from the posterior distributions for the control and variation. The most common approach for this is to use Monte Carlo simulations, which are simple, but very slow. Growth Book needs to scale to tens of thousands of metrics and thousands of simultaneous experiments and simulations are just not fast enough to give real-time results. Instead, we use two estimation methods instead: Gaussian quadratures and CLT approximations.

3.1 Which version is better?

To answer this question, we approximate a distribution D_1 which is the difference between the variations ($P_B - P_A$). The probability that the variation is better is simply $P(D_1 > 0)$, or the integral of D_1 from 0 to Infinity. Growth Book exposes this value as the "Chance to Beat Control."

For binomial metrics, D_1 is defined as: [5]

$$D_1 = P_B - P_A|X \sim N(E[P_B|X_B] - E[P_A|X_A],\ Var(P_A|X_A) + Var(P_B|X_B) \tag{3}$$

For gaussian metrics, it is defined as: [5]

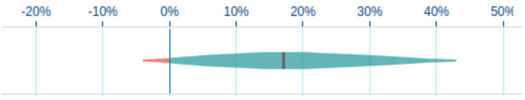
$$D_1 = \mu_B - \mu_A|X \sim N(E[\mu_B|X_B] - E[\mu_A|X_A],\ Var(\mu_A|X_A) + Var(\mu_B|X_B) \tag{4}$$

3.2 How much better is it?

To answer this question, we approximate a second distribution D_2, which is the log of relative uplift.

$$relative\ uplift = log\left(\frac{P_B}{P_A}\right) \tag{5}$$

Instead of just showing a single uplift number or a credible interval (the Bayesian equivalent to confidence intervals), Growth Book opts to show the entire distribution as a Violin Plot.



We have found this tends to lead to more accurate interpretations. For example, instead of just reading the above as “it’s 17% better”, people tend to factor in the error bars (“it’s about 17% better, but there’s a lot of uncertainty still”).

For binomial metrics, D_2 is defined as: [5]

$$D_2 = ln\left(\frac{P_B}{P_A}\right)|X \sim N(E[ln\, P_B|X_B] - E[ln\, P_A|X_A],\ Var(ln\, P_A|X_A) + Var(ln\, P_B|X_B) \qquad (6)$$

For gaussian metrics, we use the Delta method to approximate D_2 : [5]

$$D_2 = ln\left(\frac{\mu_B}{\mu_A}\right)|X \sim N\left(ln\, E[\mu_B|X_B] - ln\, E[\mu_A|X_A],\ \frac{Var(\mu_A|X_A)}{E^2[\mu_A|X_A]} + \frac{Var(\mu_B|X_B)}{E^2[\mu_B|X_B]}\right) \qquad (7)$$

3.3 Do I have enough data to call the test now?

Knowing when to stop an experiment is hard. There are usually many external factors that go into the decision. For example, if the new variation is going to save you maintenance costs down the road, you may be willing to take a small hit in conversions. Or if your experiment happens to coincide with a holiday weekend, you may want to wait longer to make sure the effects you are seeing hold up over a more normal time period. Because of these external factors, it’s not possible to completely turn the decision making over to an algorithm.

Instead of answering this question directly with a “yes” or “no”, we answer a related question using Bayesian Risk (or Expected Loss). “If I choose B and it’s actually worse, how many conversions am I expected to lose?”. This lets the human decision maker weigh all of the external factors together with the Risk to determine the stopping point of the experiment.

The simplified risk metric is written as: [5]

$$R_B = \xi_A + \xi_B - E[P_B] \qquad (8)$$

To estimate the integral ξ_A, we use Gaussian quadratures (GQ). This produces results just as good as Monte-Carlo simulation in a fraction of the time (using about 20 nodes instead of thousands) [5][7].

4 Data Quality Checks

In addition to answering the common A/B test questions, Growth Book performs data quality checks to ensure the statistical inferences are valid and ready for interpretation. We currently run two checks: Sample Ratio Mismatch (SRM) and minimum data thresholds.

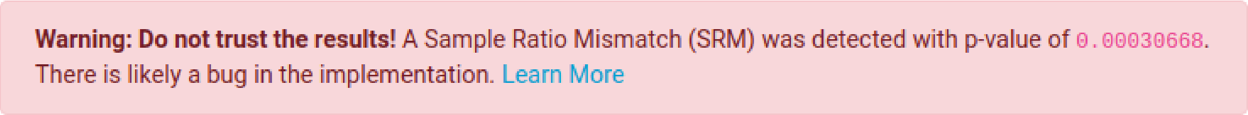
4.1 Sample Ratio Mismatch (SRM)

In addition to answering the questions above, a proper A/B testing stats engine must detect potential problems in the collected data. One of the most common issues is a Sample Ratio Mismatch (SRM). This occurs when the observed traffic split between experiment variations does not match the expected split. For example, you are expecting a 50/50 split, but observe a 48/52 split.

Growth Book calculates a simple Chi-Squared statistic to compare the observed sample sizes with the expected sample sizes for each variation:

$$\chi^2 = \sum_{i=1}^{n} \frac{(Oi - E_i)^2}{E_i} \tag{9}$$

Then we obtain a p-value and set a low threshold of 0.001 (99.9% confidence level) to avoid false positives. When a p-value below this threshold is encountered, it is shown above the results as a warning:



Warning: Do not trust the results! A Sample Ratio Mismatch (SRM) was detected with p-value of `0.00030668`. There is likely a bug in the implementation. Learn More

In a meta-analysis, Microsoft researchers found that 6% of experiments had an SRM [8]. The SRM is a symptom of a problem and not the cause. To help discover the root cause, Growth Book shows within the app a comprehensive taxonomy of problems based on that same Microsoft research study. [8]

4.2 Minimum Data Thresholds

The statistics engine just runs mathematical equations on the data it is given. It will happily tell you the "chance to beat control" at the start of your experiment when you only have 2 conversions vs 1 conversion even though it's way too early to even start thinking about results. To avoid confusion and prevent drawing false conclusions, we hide statistical inferences for a metric until a minimum threshold is reached. Both of the following must be true for Growth Book to show results for a metric:

1. Both the variation and control must have at least 25 conversions
2. At least one of them must have at least 150 conversions

If a metric has not reached this threshold yet, Growth Book will show an estimated time remaining based on the rate of data collected so far (if any) and how long the experiment has been running for.

5 Conclusion

Growth Book utilizes a combination of Bayesian statistics, fast estimation techniques, and data quality checks to robustly analyze A/B tests at scale and provide intuitive results to decision makers. The implementation is fully open source under an MIT license and available on GitHub. Our statistics engine is already top of class, but we plan to continue improving it by adding support for Informative Priors based on historical data, more data quality checks, user customization, and a wider selection of specialized metric distributions.

References

[1] Evan Miller, “How Not to Run an A/B Test”, 18 April 2010, https://www.evanmiller.org/how-not-to-run-an-ab-test.html

[2] “Misuse of p-values”, Wikipedia, Wikimedia Foundation, 31 July 2021, https://en.wikipedia.org/wiki/Misuse_of_p-values

[3] Itamar Faran, “Why You Should Switch to Bayesian A/B Testing”, Towards Data Science, 7 June 2021, https://towardsdatascience.com/why-you-should-switch-to-bayesian-a-b-testing-364557e0af1a

[4] “Prior probability”, Wikipedia, Wikimedia Foundation, 31 July 2021, https://en.wikipedia.org/wiki/Prior_probability

[5] Itamar Faran, “How to do Bayesian A/B Testing at Scale”, Towards Data Science, 21 June 2021, https://towardsdatascience.com/how-to-do-bayesian-a-b-testing-fast-41ee00d55be8

[6] “Central limit theorem”, Wikipedia, Wikimedia Foundation, 31 July 2021, https://en.wikipedia.org/wiki/Central_limit_theorem

[7] “Gaussian quadrature”, Wikipedia, Wikimedia Foundation, 31 July 2021, https://en.wikipedia.org/wiki/Gaussian_quadrature

[8] A. Fabijan et al., “Diagnosing Sample Ratio Mismatch in Online Controlled Experiments,” in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining – KDD ’19, 2019, pp. 2156–2164.