Bayesian A/B Testing

Murray Cantor, Ph.D. IBM Distinguished Engineer (Ret.)

Introduction

I asked ChatGPT-3, "What is A/B testing?". It gave an answer that is clearer than anything I might have written:

A/B testing, also known as split testing or bucket testing, is a research methodology used to evaluate the performance of two or more variants of a digital element, such as a webpage, email campaign, advertisement, or application interface. The goal of A/B testing is to identify which version generates better results in terms of specific objectives, like user engagement, conversion rates, or click-through rates.

During an A/B test, the target audience is randomly divided into separate groups, with each group being exposed to one of the variations (e.g., version A or version B). The effectiveness of each version is then measured based on a predetermined metric, such as sign-ups, sales, or time spent on the page. Once sufficient data has been gathered, statistical analysis is conducted to determine if there is a significant difference in performance between the tested variations. The variant that demonstrates superior performance is ultimately chosen for wider implementation.

A/B testing is commonly employed in digital marketing, user experience design, and product development, as it allows businesses to make data-driven decisions and continuously refine their offerings.

I then asked the chatbot, "What is Bayesian A/B testing?" Its answer, while correct, was less helpful:

Bayesian A/B testing is a variation of the traditional A/B testing approach that utilizes Bayesian statistics to estimate the performance of different versions of a digital element, such as a webpage, email campaign, or advertisement. While both traditional A/B testing (also known as frequentist A/B testing) and Bayesian A/B testing have the same goal—to determine which variation performs better—the Bayesian method offers some advantages in terms of analysis and decision-making.

Bayesian A/B testing incorporates prior knowledge and observed data to calculate the probability of each variant being the best performer. It does so by updating the prior probability with new data, which results in a posterior probability. Bayesian statistics allows for a more flexible and intuitive understanding of the results, making it easier to incorporate domain knowledge and expert opinions.

Advantages of Bayesian A/B testing over frequentist A/B testing include:

- *Interpretability: Bayesian results are more intuitive and easier to understand since they provide direct probability statements about the performance of each variant.*

- *Flexibility: Bayesian methods can accommodate changes during the testing period, such as adding or modifying new variants.*
- *Faster results: Bayesian methods often provide meaningful insights with smaller sample sizes, allowing for quicker decision-making.*
- *Better handling of uncertainty: Bayesian methods provide a more comprehensive view of uncertainty by using probability distributions, which can help guide decision-making more nuancedly.*

Overall, Bayesian A/B testing is a powerful approach for comparing different versions of digital content and making data-driven decisions. It offers a more flexible and interpretable way to understand test results and optimize digital assets based on performance.

This paper elaborates on the Chat-GPT explanation with a mathematically complete explanation of Bayesian A/B testing. It goes on to describe how to do it in practice and, with examples, flesh out the claimed advantages.

I assume the reader is familiar with (or at least aware of) random variables, probability density functions (PDFs), conditional probability, and Bayes theorem. As such, this paper is meant to be useful to practicing data scientists interested in bringing these techniques to their enterprises. It should also be helpful for budding data scientists to build out their skills.

A/B Testing

A/B testing is used to test the effectiveness of one sort of 'treatment.' For example, when testing a drug, you would randomly select two samples from the same population and give one sample the drug to be tested and the other a placebo. You measure the recovery rates of both populations to see if there is a significant difference. If there is, you deem the drug effective.

Traditionally, null hypotheses frequentist statistical significance testing is used ($p < 0.05$) to specify effectiveness. For various reasons, that approach is being supplanted by Bayesian methods. A great discussion of the weakness and the ugly history behind frequentist reasoning see (Clayton, 2021).

These days, A/B testing is often used for testing variants of web pages. In this case, two variants of the pages are presented to the population hoping for some action. Examples of a 'success' could be clicking through to a more detailed page, adding the described product to the shopping cart, or completing the sale.

E-commerce economics are very different from drug testing. Approving a drug is far more serious than choosing which web page variant to use. Hence the drug test decision requires more certainty than the web page decision.

Why Bayes?

You might choose Bayesian methods over frequentist methods for A/B testing for several reasons. Primary reasons include:

1. Small sample sizes: Bayesian methods can often provide more reliable and stable results with small sample sizes because they rely on the entire posterior distribution rather than point estimates like frequentist methods. This can be especially useful for A/B testing when you have limited data and must make quick decisions.
2. Robustness: Bayesian methods can be more robust to violations of model assumptions compared to frequentist methods, as they are based on probability distributions instead of point estimates
3. Quantification: Bayes quantifies the probability that one variation is better than the other, which is more informative for decision-making.
4. Interpretability: Bayesian methods provide probability distributions for parameters, which are more interpretable than frequentist p-values or confidence intervals.

In addition, there are further reasons which apply to more advanced applications.

1. Multiple testing adjustments: Bayesian methods can account for multiple comparisons more naturally than frequentist methods. They don't suffer from the same inflation of false positive rates associated with multiple hypothesis testing and don't require strict adjustments for multiple comparisons.
2. Modeling flexibility: Bayesian methods offer more modeling flexibility compared to frequentist methods. Using Bayesian techniques, you can more easily include covariates, hierarchical models, or other complex relationships between variables.
3. Robustness: Bayesian methods can be more robust to violations of model assumptions than frequentist methods, as they are based on probability distributions instead of point estimates.

The Mathematical Formulation

Basic A/B testing entails simultaneously running two Bernoulli trial experiments, one for asset A and one for asset B. Both experiments have the same notion of a successful trial. The experiments are run repeatedly to get several runs. The number of trials and successes for each experiment is accumulated over the runs. So, the data for each experiment is the sequence are:

$$Data_A = ((SA_r, TA_r))_{r=1}^{runs}$$

$$Data_B = ((SB_r, TB_e))_{r=1}^{runs}$$

SX_r is the number of successes, and TX_r is the number of trials for X = A, B. Since the data are accumulated across runs, the S and T sequences are increasing.

There are some essential assumptions for what follows.

- The subjects for the trials are randomly chosen from the same population.

- The trials are independent.

In frequentist A/B testing, the success propensity for A would be

$$p_A = \lim_{r \to \infty} \frac{SA_r}{TA_r}$$

And p_B is defined similarly. In practice, one continues the experiment until one of several statistical tests, such as the chi-square or binomial test, supports rejecting the null hypothesis that one of the propensities is not sufficiently greater to support choosing A or B.

The Bayesian approach treats the propensities as random variables rather than approximated constants. The benefits of Bayesian A/B testing listed by ChatGPT testing follow from taking this tack.

Bayesian A/B testing can then be summarized as follows:

- Set the stopping criterion of the form, "It is *x%* likely that the absolute difference of p_A and p_B is greater than or equal to a target difference.
- Run the experiments to capture the $Data_A$ and $Data_B$ after each run.
- After each run, use Bayesian parameter learning to learn the PDFs of p_A and p_B.
- Inspect the expected values of p_A and p_B to see which one is greater.
- Compute the random variable $|p_A - p_B|$ and compute its probability of being above *d*.
- Track the trend of the probabilities to see if they are trending to *X*.

Learning the PDFs of p_A and p_B

For A/B testing, we know the S and T for each experiment, and we want to find each experiment's biases, p_A and p_B.

Using Bayes Theorem

This section covers how to compute the biases from the data. From the frequentist definition of the bias, p is a scaler in the interval [0, 1]; if p = 0, every trial will fail; if p = 1, every trial will succeed. The higher the p, the more likely the success of a trial.

With the assumption that the trials are independent, the likelihood of having *S* successes with *T* trials with bias *p* is given by the Binomial distribution (Equation 1). Using the language of random variables.

$$P\big((S,T)\big|p\big) = \binom{T}{S} p^S (1-p)^{T-S}$$
$$\binom{T}{S} \textit{ is } T \textit{ choose } S$$

Equation 1 The Binomial Distribution

We use the convention $0^0 = 1$.

Equation 1 gives *P((S, T)|p). We need P(p|(S, T))*. Going from one to the other is the point of Bayes theorem:

$$PDF(p) = P(p|(S,T)) = \frac{P\big((S,T)\big|p\big)P(p)}{P(S,T)}$$

Equation 2 Bayes theorem

Equation 2 gives the formula for finding the PDF of the bias. In our context, the terms of the equation are:

- *P(p|(S, T))* is called the 'posterior.'
- *P(p)* is called the 'prior' belief, the initial PDF of the parameter without accounting for evidence. This will be discussed later.
- *(S, T)* is the observed data of successes and trials.
- *P((S, T)|p)* is the likelihood of the data for any $p \in [0,1]$.
- *P(S, T)* is called the marginal. Finding this turns out to be simple.

Choosing The Prior

The first step in any Bayesian calculation is to choose the prior. For A/B testing, it is reasonable to assume that at the onset, there is no basis for assuming an expected value of p. That is, one should assume that all values of p are equally likely. Since the support of the parameter is the closed interval,[0, 1], we would set *P(p) = 1 for all p* in [0,1]. This is called a 'uniform prior.' Using a uniform prior is sometimes called 'the principle of indifference.' It is the best choice to avoid prejudices.

Since the result of Bayes equation is a PDF, its integral of the domain must equal 1. This allows one to deal with the marginal. Thus, the pseudo-code for computing the PDF is:

Given data (S, T) of successes and trials:

1. Discretize [0,1] to get an array $D= [0, d_1, d_2, \ldots, 1]$ of length n.
2. Compute the likelihood array $L = [P((S,T)|d_i)]$ for each i.
3. Numerically integrate L over array D to get the total area, A.
4. Normalize L by dividing by A.
5. Interpolate the normalized L to get the PDF of p.

Figure 1 shows the calculation output for different choices of S and T. Note that the choices of S and T in the four examples have roughly the same ratio.

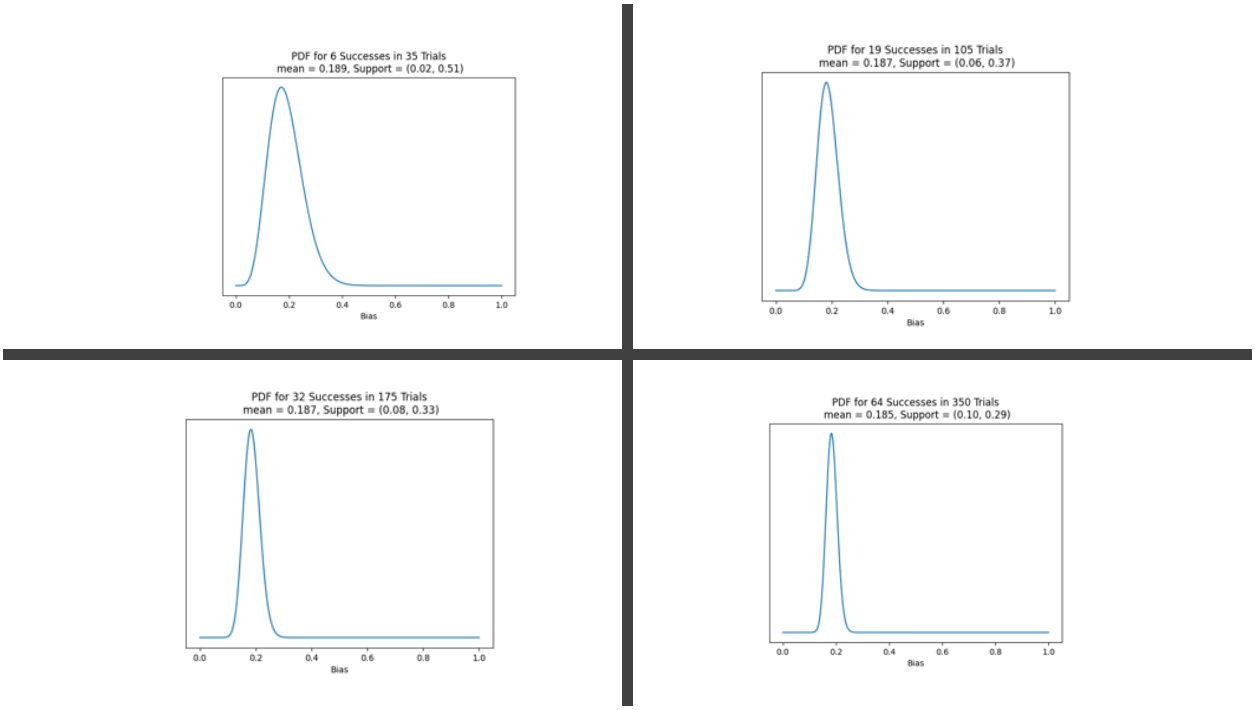


Figure 1. The PDFs of p for various choices of S and T

In Figure 1, the supports are the 100% error bounds. That is, the parameter is 100% likely to fall in the support interval. Note that the means of the four examples are close and that the support narrows rapidly with the number of trials.

Visualizing the Results

Having the PDFs of the parameters is the basis of Chat-GPT's assertions of the Bayesian 'transparency' and 'deeper insights' advantages. One can generate parametric and non-parameter statistics (e.g., percentiles) with the PDFs. One can even generate random samples for use in random variable calculations. As discussed below, they can determine whether the stopping criterion is met. In addition, the samples can be used to compute the choice's economics.

A dual PDF diagram is an excellent way to understand whether A or B is better (Figure 2).

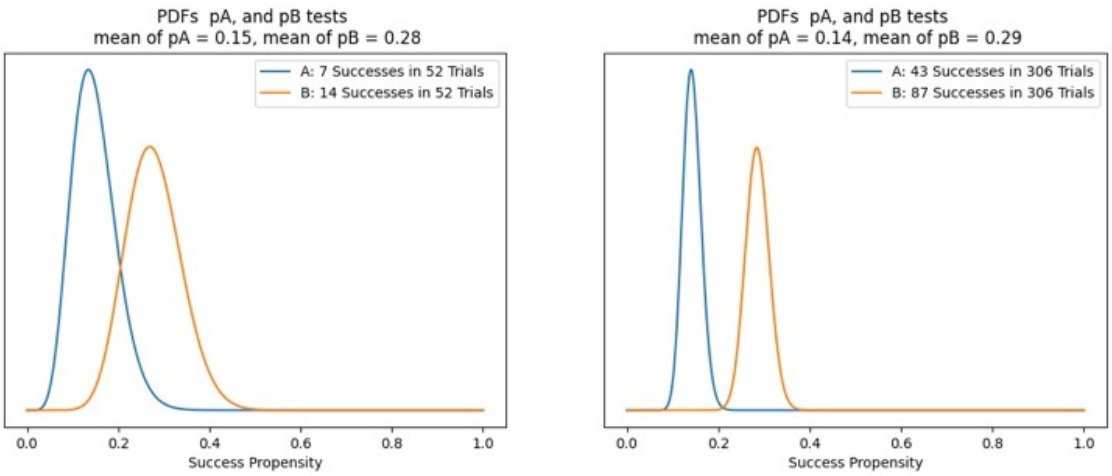


Figure 2 A pair of dual PDF graphs

The graphs are overlayed PDF diagrams. The x-axis is the propensity, and the height of the curves is the probability of x. The left-hand graph uses a small subset of the data used for the right-hand chart. The means of the PDFs are roughly the same in both graphs. However, the PDFs are wide on the left, with much overlap. They narrow and separate with more data.

Figure 2 makes it clear that B is better than A. But is it different enough to meet the stopping criteria? Let's suppose that we want to be sure that it is 85% certain that the propensity of B is at the p_B is at least .1 above the propensity of p_A.

In practice, this means if we take a random sample of p_B and of p_A, the difference is 85% likely to be greater than 0.1. This can't be seen in Figure 2. Even if the PDFs don't overlap, they still might not be the criterion.

A way to see if the criterion is met is to compute the random variable of the difference. There are various ways to do this. These days the most efficient way is to use Monte Carlo simulation.

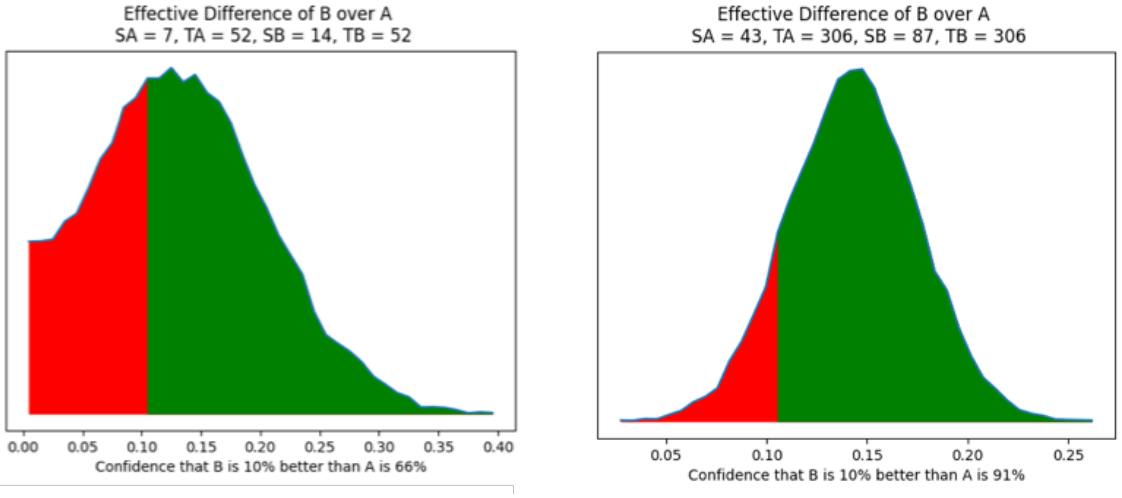


Figure 3 Criterion charts for the PDFs in Figure 2

Each of the graphs in Figure 3 is a PDF of the absolute difference of p_A and p_B. They are computed from the PDFs in Figure 2. The x-axis is the absolute difference between the PDFs. The region above the target difference in the x-axis is green. The area of the green part is the probability of the difference meeting or exceeding the target. That area is documented in x-label.

When to Stop

With the above, we can recast the A/B testing as follows:

While simultaneously running the experiments,

- Periodically assemble the successes and trials for each experiment (SA, TA) and (SB, TB).
- Compute the PDFs of the pA and pB parameters for each experiment.
- Set a different target and confidence level. For example, if you were 80% confident that the pB is at least 0.1 better than A, you would choose treatment B.
- The testing is stopped when the confidence in the target difference is reached, or it becomes clear it won't be reached.

The choices of the target and confidence level depend on the economics of the outcome of the experiments.

In addition to Figure 2 and Figure 3, one more chart is needed to figure out when to stop.

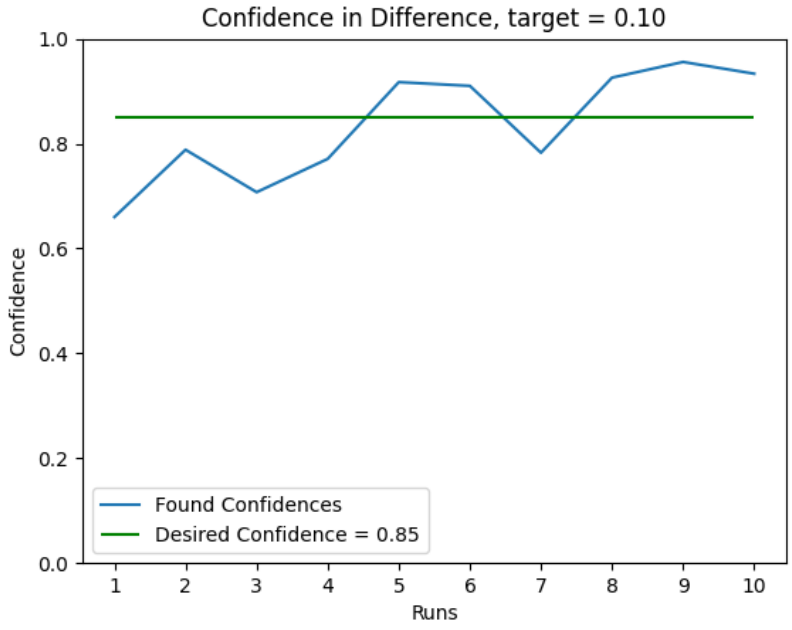


Figure 4 The confidence trend chart

Figure 4 shows the area of the green region for each run of the experiment. This chart raises an important caveat. A stopping criterion is met when the curve exceeds the desired confidence line. The confidence may not be monotonic. You may want to have several more runs before committing to a choice. Or, to be more conservation, apply regression analysis to the trend curve to get error bounds.

To show how the stopping criteria work in practice, here are three examples:

1. The difference between B over A is well above the desired target.
2. B is above A, but the difference is below the target
3. B is above A are have similar propensities

These examples were built using an A/B test simulator.

Simulation 1

	A		B	
RUN	Successes	Trials	Successes	Trials
1	5	40	9	40
2	15	88	27	88
3	21	138	38	138
4	28	182	58	182
5	33	223	76	223
6	39	274	90	274
7	47	324	105	324
8	56	380	122	380
9	65	433	135	433
10	74	490	149	490

Table 1 The data for Simulation 1

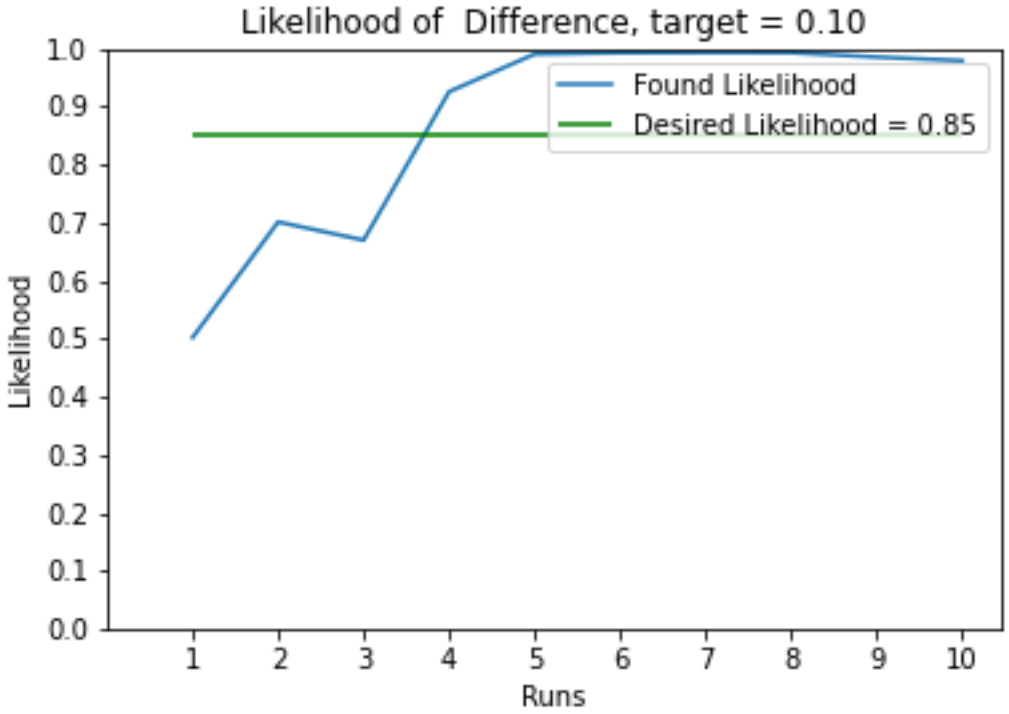


Figure 5 The confidence chart for simulation1

Note that by run 4, and certainly by run 5, the propensities are clearly different enough to choose one over the other.

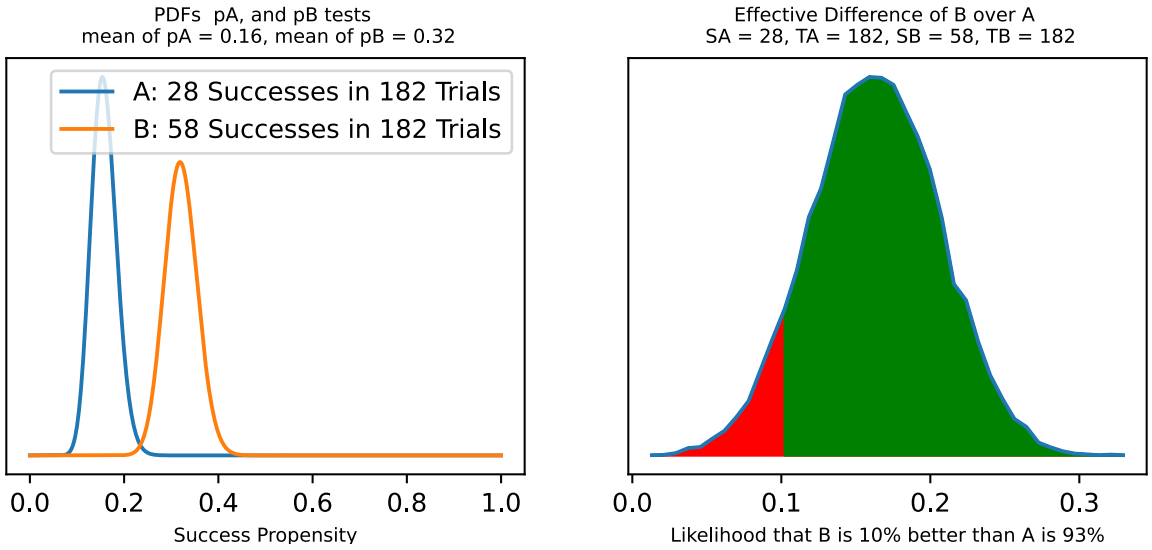


Figure 6 The dual PDF and criterion charts for run 4

Note that the PDFs barely overlap, and you could choose B over A in just 4 runs. In this case, one could stop the experiment at run 2 or 3.

Simulation 2

	A		B	
RUN	Successes	Trials	Successes	Trials
1	6	54	12	54
2	22	109	22	109
3	25	149	30	149
4	35	200	47	200
5	43	251	62	251
6	50	295	72	295
7	65	348	83	348
8	70	395	96	395
9	78	445	110	445
10	90	497	120	497

Table 2 The data for simulation 2

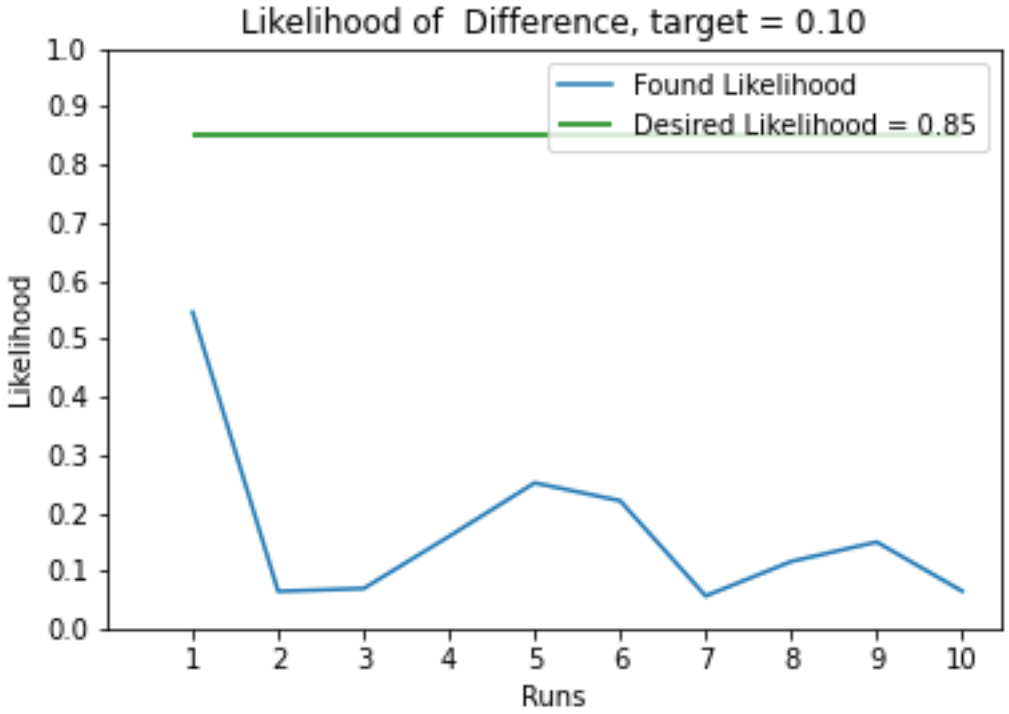
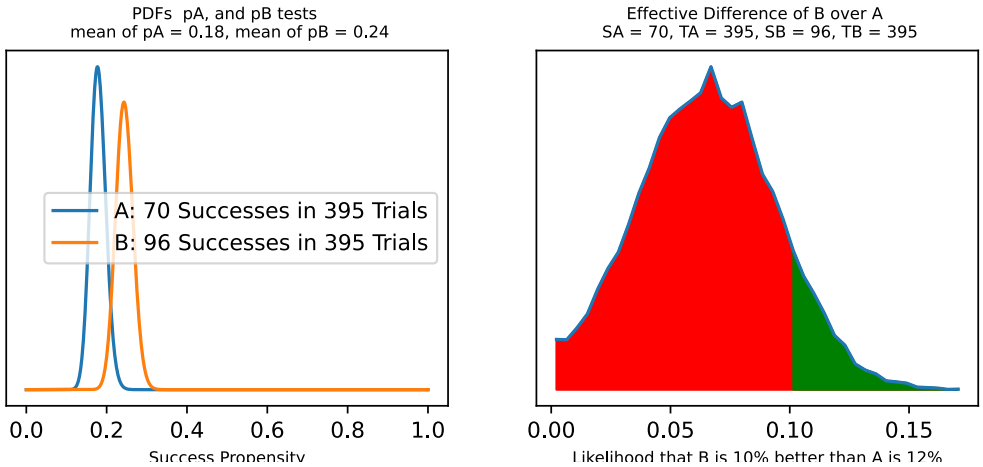


Figure 7 The confidence chart for simulation 2

Here is the dual chart for run 8. Note that A and B overlap to an extent. There is little reason to choose one over the other. The difference chart reinforces this. From the confidence chart, it is reasonable to end the experiment around run 5.



When B is closer to A, it is unsurprising that it takes more runs to be confident. Here is the confidence trend chart.

Simulation 3

	A		B	
RUN	Successes	Trials	Successes	Trials
1	5	46	6	46
2	21	109	18	109
3	33	162	26	162
4	44	211	29	211
5	55	254	34	254
6	68	304	44	304
7	74	348	49	348
8	85	401	57	401
9	95	451	62	451
10	101	495	69	495

Table 3 Data for when B is less desired difference above A.

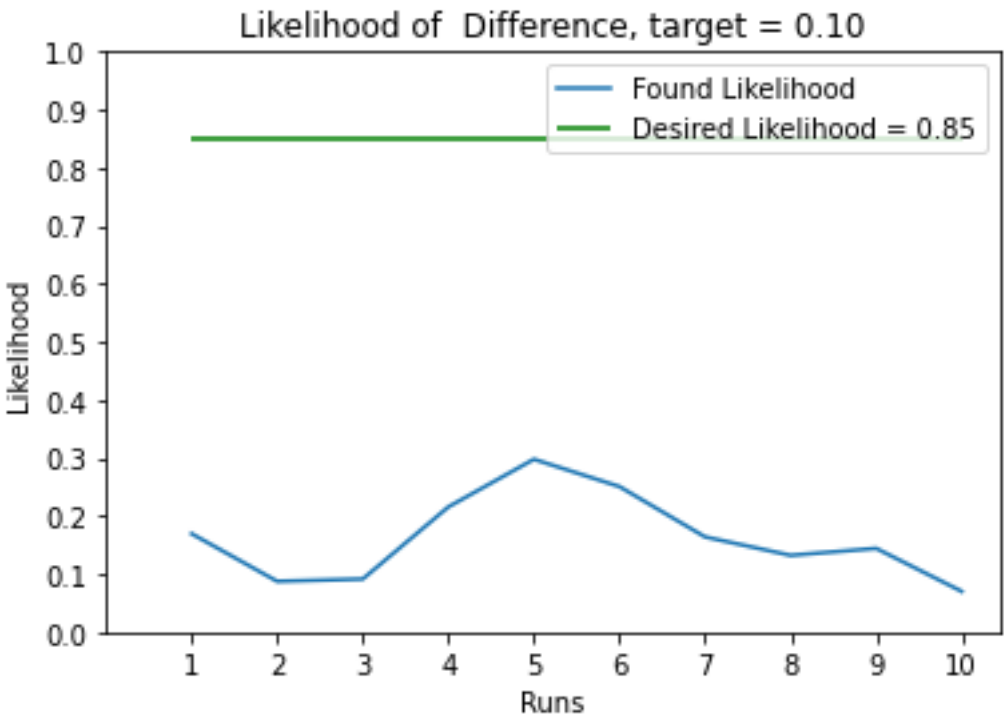


Figure 8 The confidence trend chart for Table 3

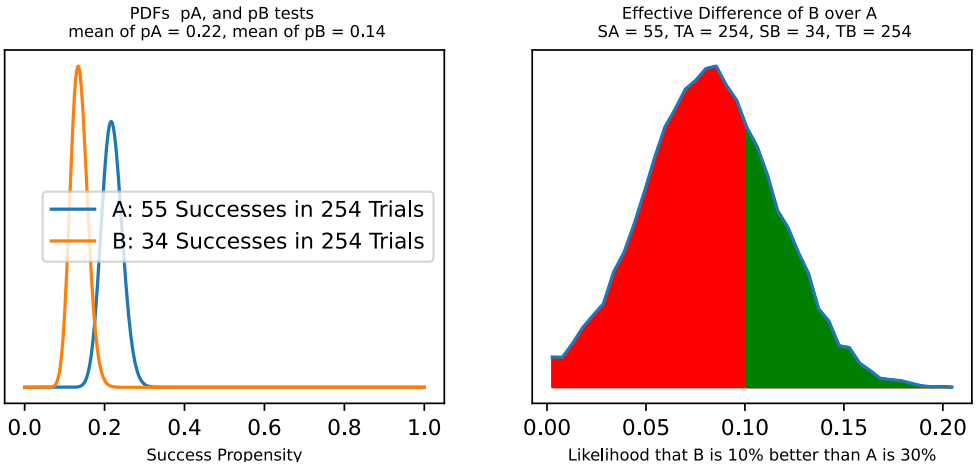


Figure 9 The dual PDF and criterion charts for run 5

At the first run (Figure 9), we see the means are close with much overlap. The PDFs are essentially the same. With more runs (Figure 10), the difference between the PDFs comes into focus. With this focus, it is more evident that the choice criterion will be met.

Figure 10 The dual PDF and criterion charts for run 10

With this data, it is reasonable to call off the experiment at run 4 or 5.

More Benefits of the Bayesian Method.

Note that Chat-GPT benefits are supported by use direct use and visualization of the PDFs. Other benefits include:

- The math is more straightforward and more robust.
- It is more accurate and reliable for hypothesis testing than frequentist methods, which can give the wrong answer up to 20% of the time. Again, see (Clayton, 2021).
- Providing probability density functions (PDFs) that can be used for further analysis offers a complete understanding of the uncertainty surrounding the estimated effects of the treatment.

Overall, Bayesian A/B testing offers speed, flexibility, accuracy, and the ability to use PDFs for economic modeling, making it a valuable tool for businesses and researchers alike.

Works Cited

Clayton, A. (2021). *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science.* Columbia University Press.