Bayesian Approach in Growthbook

Team	@Andreas Koukias @Liubimova Ekaterina Vladimirovna (Deactivated) @Daniel Evangelista Vareda
Goal	Understand and Dive into Bayesian Statistics and how Growthbook manages them
Reference	• Growthbook White Paper • Growthbook Page on Statistics • Bayesian Statistics Specialization • Pymc Video • Pymc Page • Article from Pymc • White Paper - VWO • Bayesian Workflow • Bayes Theorem Video
Next Steps	• Test an A/B Test and A/A test for a same experiment using Bayesian Calculations • Dive on the Growthbook calculations and Open Code • Understand about iterations on the Priors for Growthbook

---

Guidelines and Findings

- Growthbook adopts **uninformative** priors (“We know nothing about our data. All outcomes have the same likelihood. “)
  - Outcome:
    - More **variability** in the results
    - Need for a larger sample size
    - Higher Rate of **False Positive or Negatives**
  - What then?:
    - Allow the experiment to run to a **full seasonality window**
    - Check the **MDE** of the sample size (how precise it is)

- **Chance to Beat Control** is calculated as the area under the curve of the HDI with a positive outcome

- **Uplift** is the value under the most dense part of the HDI

- Growthbook UI may be **biased toward positive decisions**
  - Outcome: **Chance to beat control** and **Uplift** may lead to wrong decisions
  - What then:
    - Check for the **Chance to Not Beat Control (100% - X%)**
    - Don't present the uplift as the final impact of the feature
- **Drawn / Lost / Win** is a business definition based on ROPE

Presentation

Sign in to access Google Drive Presentation

Bayesian Store ...e.ipynb
07 Feb 2024, 02:17 PM

Workshop - Part 1:

Sign in to access Google Drive File

Sign in to access Google Drive Document

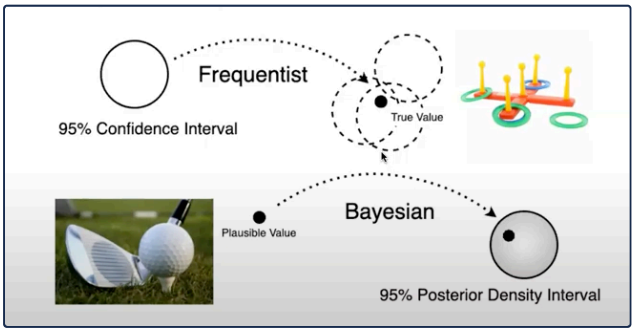
Workshop - Part 2:

Sign in to access Google Drive File

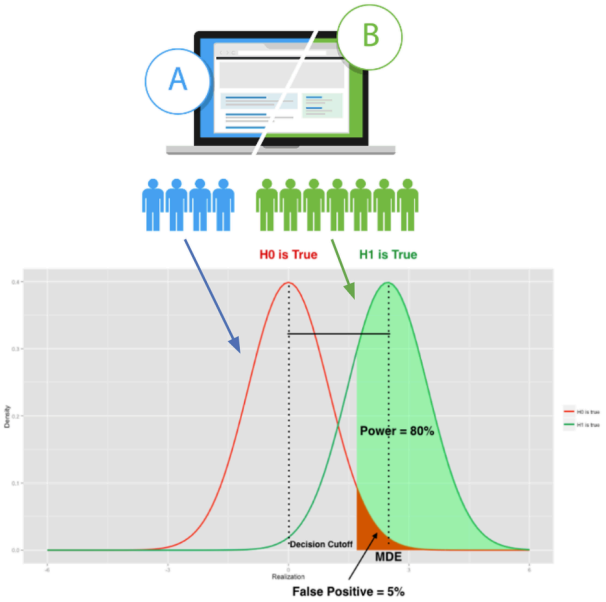
Sign in to access Google Drive Document

Frequentist vs Bayesian Approach on Experimentation

Methods



Frequentist Approach



Components:

- **Null Hypothesis:** In frequentist statistics, experiments often start with a null hypothesis, which is a statement that there is no effect or no difference between groups.
- **P-value:** The probability of obtaining an effect at least as extreme as the one in your sample data, assuming the null hypothesis is true.
  - *If your p-value is 0.05 %, it means that 5% of the time you would see a result similar to this one (or as extreme as) if your null hypothesis was true*
- **Confidence Intervals:** Range of values within which a parameter is expected to lie, with a certain degree of confidence.

Pros:

- Simplicity: Frequentist methods are often simpler and more straightforward to apply.
- Widely Accepted Standards: Many scientific fields and regulatory bodies have established standards based on frequentist statistics.
- No Need for Prior Knowledge: Frequentist methods do not require prior distributions, which can be advantageous when no prior information is available.

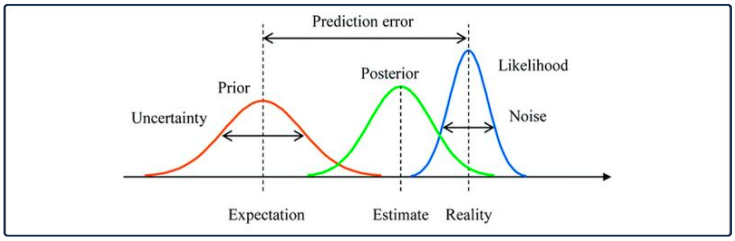
Cons:

- P-value Misinterpretation: P-values can be misinterpreted; they don't tell you the probability that the hypothesis is true.
- Confidence Interval Misunderstandings: Confidence intervals are often misunderstood as the probability of the parameter lying within the interval.
- Reliance on Large Sample Sizes: The accuracy of frequentist methods improves with larger sample sizes.

Bayesian Approach

Components:

- Prior Distribution: A probability distribution representing knowledge or uncertainty about a parameter before considering the current data.
- Likelihood: The probability of the observed data under different parameter values.
- Posterior Distribution: Updated beliefs about the parameter after observing data, derived from the prior distribution and the likelihood.



Pros:

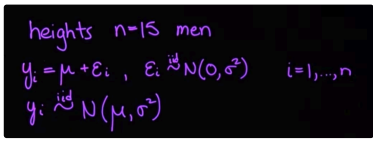
- Incorporates Prior Information: Bayesian methods allow the integration of prior knowledge or beliefs into the analysis.
- Probabilistic Interpretation: Provides a probabilistic interpretation of results, such as the probability that a parameter lies within a certain range.
- Flexibility: Can be more flexible in dealing with complex models and smaller sample sizes.

Cons:

- Complexity: Bayesian methods can be computationally intensive and complex, especially for large or complicated data sets.
- Subjectivity of Priors: The choice of prior can be subjective and may significantly influence results.
- Less Standardized: Less standardized in some fields, leading to potential issues in acceptance or interpretation.

Example 1

Consider the heights of 15 people. The individual height(yi) can be defined as the mean of the height (mu - u) plus an individual error (Ei), with this error being part of a normal distribution (N), with average 0 and variance (Sigma). We can also write the individual heights in function of the Normal Distribution as seen below.



The frequentist approach considers mu and sigma fixed and unknown. To assert the uncertain, it would consider how much this variables will change, the frequentist approach calculates how much these results would variate if we were to repeat the experiment over and over again with 15 people.

For the bayesian approach, treats the mu and sigma as variables with their own distribution (priors).

The likelihood is defined as how the data probability behaves, like the density of a variable.

Likelihood = P(y | T)

Prior = P (T)

Posterior = P (T | y)

Once knowing these components, we can then simulate more data with similar distributions, simulating drafts and experiments.

Example 2

Your brother or sister has a coin that you know is loaded and comes head 60% of the time.

He then comes up to you, with a coin, you are not sure if loaded or not, and he want to beat money it is gonna come up heads.

Because you are unsure if the coin is fair, he allows you to flip it five times. This gives you two heads and three tails.

- Which coin is it?

When we compute the likelihood in this problem (Frequentist Method), given that X=2 (heads), the chances of the coin being loaded is 0.1323.

Under the Bayesian approach, you can input prior information:

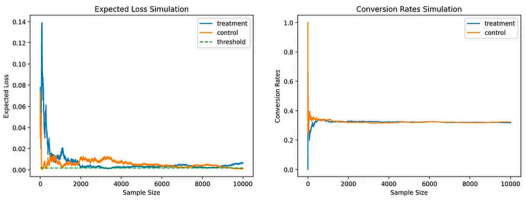
Considering you know your brother, you think there is a higher chance that the coin is loaded (60%). By adding the prior to your calculations, you get a new probability that the coin is fair: 0.388.

Peeking

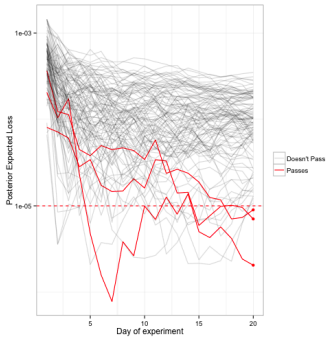
Both methods are subject to errors related to peeking (seeing the data every day and taking action upon it). On Frequentist approach this error is more relevant, while on Bayesian, the constant update of data reduces the issue.

For example, if we ran an experiment as the picture below:

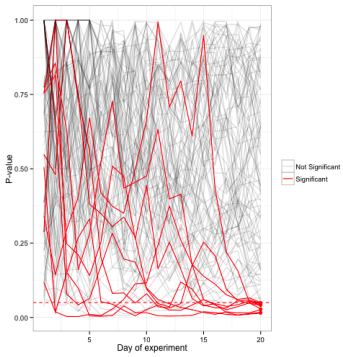
If we start the data collection and monitoring, we will see that, over time, the values might differ / oscilate a lot before stabilising. If after a few samples collected we take action upon it, we will be very likelly be taking a decision based on a false positive or negative results.



Collecting samples for a bayesian (left) and a frequentist (right) approach



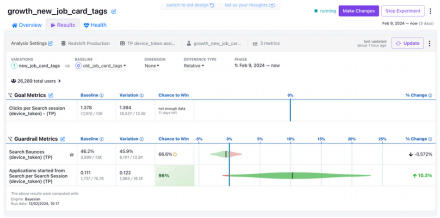
False positive rate over time for a bayesian experiment.



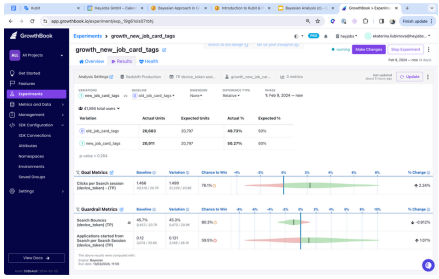
False positive rate over time for a frequentist experiment.

As Growthbook uses Uninformative Prior, the issues with peeking, mainly when the sample size is smaller or the seasonality period has not been completed, is that **results will vary a lot and may lead to wrong conclusions. It takes time and a lot of iterations for the algorithm to find the more likely distribution.**

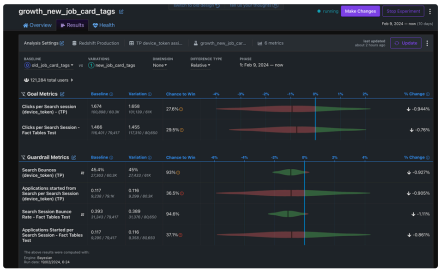
See below an example of an AB test where the metrics change drastically. Although the credible interval for the main metric always contains zero, the trend (based on the region with more concentrated data - the charts belly) shifts completely from positive to negative. The same happens to the application started metric: It goes from a 96% chance to beat control to a 37% one with a high risk involved.



This screenshot shows the experiment after 3 days.



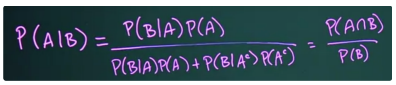
In the following day, the application started metric has already changed. The main metric (clicks per session) has a positive trend but it still contains zero.



Finally, there we can see that the main metric (clicks per session) has shifted to a negative trend, still containing zero. The same has happened to applications started.

Bayes Theorem

Bayes' Theorem is a fundamental concept in probability theory and statistics, particularly in Bayesian statistics. It describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Here's a detailed explanation:



Bayes' Theorem is mathematically stated as:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

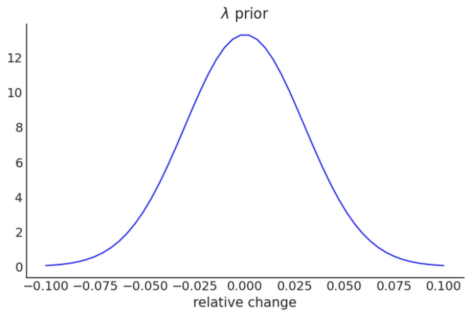
Where:

- $P(A|B)$ is the **posterior probability:** the probability of event A occurring given that B is true.
- $P(B|A)$ is the **likelihood**: the probability of observing B given that A is true.
- $P(A)$ is the **prior probability**: the initial probability of A before considering the evidence B.
- $P(B)$ is the **marginal probability**: the total probability of observing B.

Components

- **Prior Probability($P(A)$):** This represents our initial belief about the probability of an event A before we have any additional information.A prior is your uneducated opinion - it's what you believe before you have evidence. For example, if I believe all possible values of conversions are equally likely, I might choose the function P(λ) = 1.0.
- You are in the preparation to run an experiment B vs holdout A. You might be interested in increasing the mean of statistics (average bill)

- Do you expect you have a 1000% increase? Very sure No
- Do you expect you have a 100% increase? Very sure No
- Do you expect you have a 10% increase? Unlikely
- Do you expect you have a 3% increase? Maybe
- Do you expect you have a 3% decrease? Maybe



- **Likelihood (*P*(*B*|*A*)):** This is the probability of observing the evidence *B* given that *A* is true. It's how we incorporate the new evidence into our beliefs.
- **Marginal Probability (*P*(*B*)):** Often the most challenging part to calculate, this is the probability of the observed evidence under all possible outcomes. It's essentially a normalizing constant to ensure that the probabilities sum up to 1.
- **Posterior Probability (*P*(*A*|*B*)):** This is what we want to find out. It's our updated belief about the probability of *A* after taking evidence *B* into account

$$P(\lambda|\text{evidence}) = \frac{P(\text{evidence}|\lambda)P(\lambda)}{P(\text{evidence})}$$

The left side of the equation represents your posterior - your opinion after observing the evidence (tested). For the right side of the equation, the expression P(λ) represents your prior - your opinion before observing any evidence. The expression P(evidence|λ) represents the probability of observing the evidence assuming you know the true value of λ. Finally, the term P(evidence) is the probability of having observed the evidence you did observe.

Results

- **High-Density Interval:** Represents a range of the most probable values (range X to Y represent the most probable effect)
- **Bayesian Credible Intervals**: The posterior probability that the parameter is in a 95% credible interval. Unlike the confidence interval in frequentist statistics, which is based on hypothetical repeated sampling and is associated with the long-run frequency of covering the true parameter value, **the credible interval in Bayesian statistics is a direct statement about the uncertainty surrounding the parameter given the observed data.** For example, saying that the 95% credible interval for a parameter is (1.2, 2.7) means that, given the data and the model, there is a 95% probability that the true value of the parameter is between 1.2 and 2.7.

Confidence vs Credible Interval

Confidence Interval (Frequentist):

A confidence interval in frequentist statistics is an interval estimate of a population parameter. For example, a 95% confidence interval for a parameter is computed from sample data and has the property that in repeated samples, 95% of such intervals would contain the true parameter.

However, it does not mean that there is a 95% probability that the specific interval computed from a particular sample contains the true parameter. This is a common misunderstanding with confidence intervals. In frequentist statistics, the true parameter is considered fixed and the confidence interval is the random variable.

Credible Interval (Bayesian):

In contrast, a credible interval in Bayesian statistics does have the interpretation that many people intuitively assign to a confidence interval. A 95% Bayesian credible interval is an interval estimate which contains the true parameter value with a probability of 95%, given the observed data.

In Bayesian statistics, the parameter is not considered fixed but rather a random variable. So when you say "there is a 95% probability that the parameter is in the credible interval", you're expressing uncertainty about the parameter value, given the observed data.

- **Region of Practical Equivalence (ROPE):** Whether the difference between two groups (e.g., control and treatment groups) is practically meaningful
  - If the bulk of the posterior distribution lies within the ROPE, it suggests that the observed difference is not practically significant. In other words, the treatment hasn't produced a meaningful change compared to the control.
  - If the posterior distribution lies outside the ROPE, it indicates that the observed difference is practically significant. This suggests that the treatment has produced a meaningful change compared to the control.
- **Trace**: In the context of Markov chain Monte Carlo (MCMC) methods used in Bayesian statistics, a trace is a series of parameter values produced by the algorithm. The trace shows how the MCMC sampling process explores the parameter space. Trace plots can be used to assess the convergence of the MCMC sampler. They show the sampled values on the y-axis and the order in which they were sampled on the x-axis.

HDI vs Credible Interval

The difference between a general credible interval and the HDI lies in how these intervals are constructed. A credible interval might not necessarily contain the most probable values, it just indicates that the parameter lies within this range with a certain probability.

On the other hand, the HDI is specifically the interval containing the most probable values. For a unimodal distribution, the HDI would contain the mode (the peak of the distribution) and would cut off the tails.

If the distribution is symmetric like a normal distribution, the 95% HDI and the 95% credible interval would be the same and could, for instance, be (1.2, 2.7). But if the distribution is skewed, the HDI and the credible interval might differ.

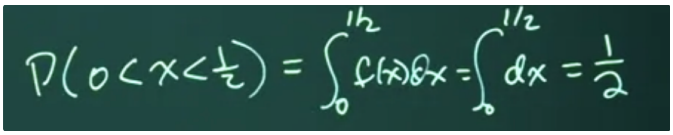
For a skewed distribution, the HDI would be shifted towards the skew. For example, if the distribution was positively skewed, the HDI might be something like (1.5, 3.0) instead of (1.2, 2.7). The exact values would depend on the shape of the distribution.

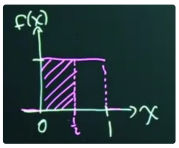
Bayesian Theory

Uniform Distribution

Definition: The uniform distribution is a type of probability distribution where all outcomes are equally likely. In a continuous uniform distribution over an interval [a,b], the probability density function is constant. The same principle applies to the discrete uniform distribution, where each distinct outcome has an equal probability.

Characteristics: In a continuous uniform distribution, the probability density is 1 / b−a for x in[a,b] and 0 otherwise. It's a way of expressing complete ignorance about the variable's value within a certain range.



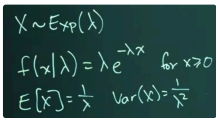


Uniform Prior: In Bayesian analysis, a uniform distribution is often used as a prior distribution. This is known as a "non-informative" or "flat" prior because it assigns equal probability to all possible values of the parameter within the specified range, indicating a lack of prior knowledge.

Implications: Using a uniform prior means that before observing any data, you have no reason to believe any value within the range is more likely than any other. It represents a state of complete ignorance or neutrality about the parameter's probability.

Exponential Distribution

Definition: The Exponential distribution is a continuous probability distribution used to model the time between events in a Poisson process. It is defined for x≥0 with a single parameter λ (rate parameter), where the probability density function (PDF) is λe (−λx) .



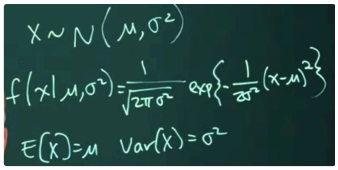
Memoryless Property: A key feature of the Exponential distribution is its memoryless property, meaning the probability of an event occurring in the future is independent of the past.

Use as a Prior: In Bayesian analysis, the Exponential distribution is often used as a prior for parameters that are positive and have a skewed distribution, such as rates (e.g., failure rates in reliability testing).

Implications: Using an Exponential prior indicates that lower values of the parameter are more likely than higher values, but without a specific upper limit.

Normal Distribution

Definition: The Gaussian distribution is a continuous probability distribution characterized by its mean μ and standard deviation σ. Its PDF is given by :



Symmetry and Bell Shape: The distribution is symmetric around the mean and has a bell-shaped curve.

Use as a Prior: The Gaussian distribution is commonly used as a prior for parameters that can take any real value, especially when there is some prior knowledge or belief about the central tendency and variability of the parameter.

Implications: A Gaussian prior assumes that values near the mean are more likely than values far from the mean, with the likelihood decreasing symmetrically in both directions.

Beta Distribution

Beta Distribution is defined as :

$$f(x;a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, 0 \le x \le 1$$

Where:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Suppose the prior P(λ) = f(λ; a, b). Suppose that the variant was displayed to n visitors and c converted. Then the posterior is given by

This theorem allows us to compute a posterior for conversion rates in a simple way, provided the prior is a beta distribution. Many possible priors can be represented via a beta distribution. A **uniform prior** is given by the choice (a, b) = (1, 1). Using the Beta distribution, we need to track only 2 numbers to compute a posterior - n and c. Then, whenever the posterior needs to be manipulated, we can compute

the posterior directly.

Practical Example

Context

Imagine you want to test 3 different campaigns. 1 is your control and the other 2 are the new variant:

1. **"Buy More, Save More!"** Implement a tiered discount system where customers receive a larger discount as they spend more, e.g., 5% off for purchases over $ 50, 10% off for purchases over $75, and 15% off for purchases over $120.
2. **"Bulk Bargains!"** Offer significant discounts on bulk purchases of non-perishable items, such as canned goods, paper products, and cleaning supplies.

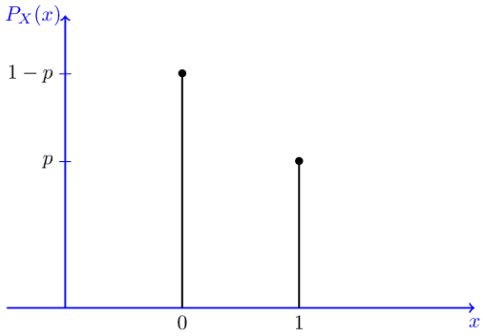
You also know that the **baseline for purchase is 97 purchases per period and the probability of a user purchasing an item is 26%.** Also**,** as we are a small company, changes are expected to vary. ±5% are good enough.

Priors

We need to define how we expect the purchase and the probability to behave prior to the experiment start.

For the **purchase probability** itself, suppose you believe this follows **a Bernoulli Distribution (0 to 1)**

$$X \sim Bernoulli(p)$$



For the **Average Purchase amount**, suppose you believe this is a **log-normal distribution** with mean and variance that follows the rule:

$$C \sim Ber(p)$$
$$Q = \begin{cases} Q \sim LN(\mu, \sigma) & C = 1 \\ 0 & C = 0 \end{cases}$$

<details>
<summary>Why Log-Normal?</summary>

A log-normal distribution is used in this context for a few reasons:

1. **Non-Negativity**: The log-normal distribution is defined only for positive values. Since purchase amounts can only be positive (you can't have a negative purchase amount), the log-normal distribution is a good choice.
2. **Skewness**: The log-normal distribution is skewed to the right, meaning it has a long tail on the right side. This can model situations where most purchases are small, but there's a long tail of larger purchases. In the context of this store, it might be that most customers only buy a small amount of nuts, but a few customers buy a very large amount.
3. **Multiplicative Effects**: The log-normal distribution is the distribution of a product of random variables. If the purchase amount is influenced by multiple factors (like the number of items bought and the price per item), and these factors are independent and randomly vary, then the total purchase amount could follow a log-normal distribution.
4. **Variability**: The log-normal distribution can model situations where there's a lot of variability in purchase amounts. Some customers might only buy a small amount of nuts, while others might buy a lot. The log-normal distribution has a wider shape that can capture this variability.
</details>

For the **Likelihood,** we can consider it a **Mixture Model**, assuming that the data comes from one of two distributions: a Dirac delta distribution centered at 0 (representing no purchase) or a log-normal distribution (representing a purchase).

We also can define the `weights` argument, which is used to determine the mixture component for each observation. It creates a two-column matrix where the first column is the probability of not making a purchase (1-p), and the second column is the probability of making a purchase (p).

Posterior

What the model does is:

1. Computes your priors for
   a. Average purchase
   b. Probability of purchase
   c. Purchase Probability Change
   d. Purchase Average Change
   e. Deviation
   f. Likelihood (Mixture Model)

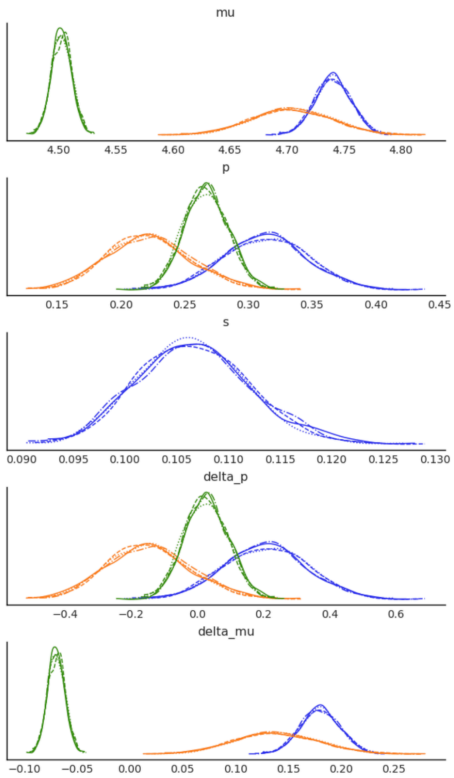
2. Computes your actual data (collected) and the posterior distribution
3. Samples from it using MCMC or any other method and define the trace
4. Collect the Posterior Distribution to define the High Density Interval

Below you can see the trace for:

- **mu**: Log of Average Purchases Amount
- **p**: Probability of Purchasing
- **s**: Data Variance
- **delta-p**: the relative difference between the observed purchase probability p and the baseline probability of 26%
- **delta-mu**: the relative difference between the observed average purchase amount mu and the baseline purchase amount of 97

Each curve represents a variation:

- **Variation 2**
- **Variation 1**
- **Control**



- **Mu**

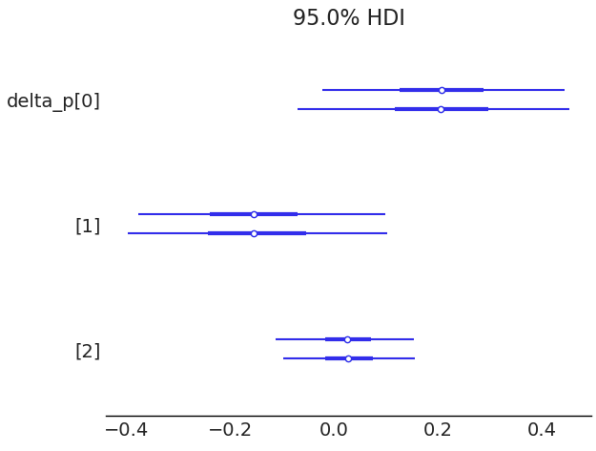
We can see that variation 1 and variation 2 have a change in mu, but and Control actually doesn't (exp(4.5) = 90 vs exp(4.7) = 109 vs 93 Baseline). We can see this effect on the **delta_mu** chart

- **p**

We can see that the probability of purchase has not changed for Control, but variation 1 and 2 have different effects. Variation 1 decreases the chance of buying while Variation 2 increases it. The same thread can be seen on the **delta_p** chart

We can see this effect also when we consider the **High-Density Interval** for each variant.

The blue line is the trace generated by sampling and the x-axis is the delta_p value.



Growthbook

How it operates

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

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 and the total number of users , we can update the Prior to get our Posterior distribution.

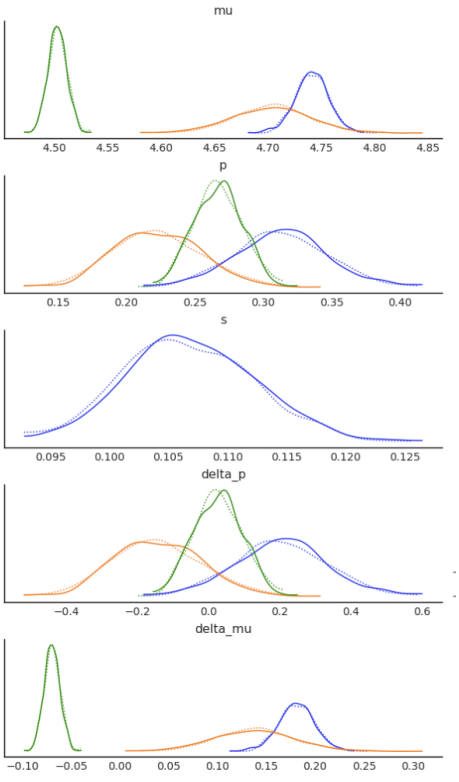
Gaussian Metrics

For count, duration, and revenue metrics, we use a Gaussian (or Normal) Prior

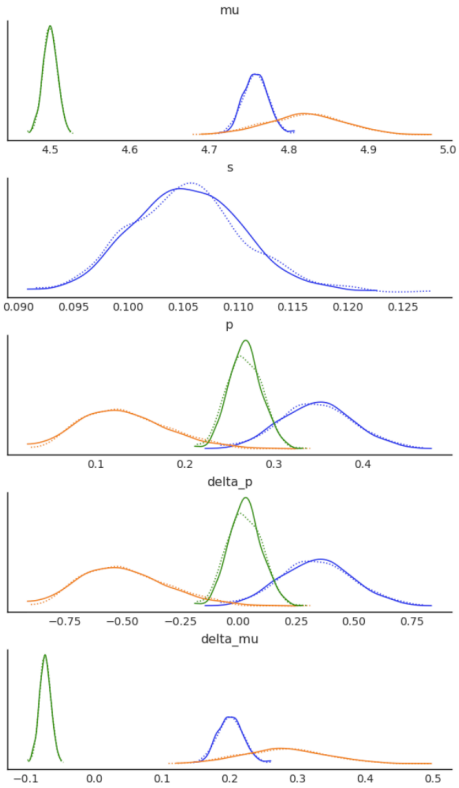
But what is the impact? It might take more time to find the actual values for the experiment:

Compare the results for the example above when we use uninformative priors:

Informative Priors



Uninformative Priors

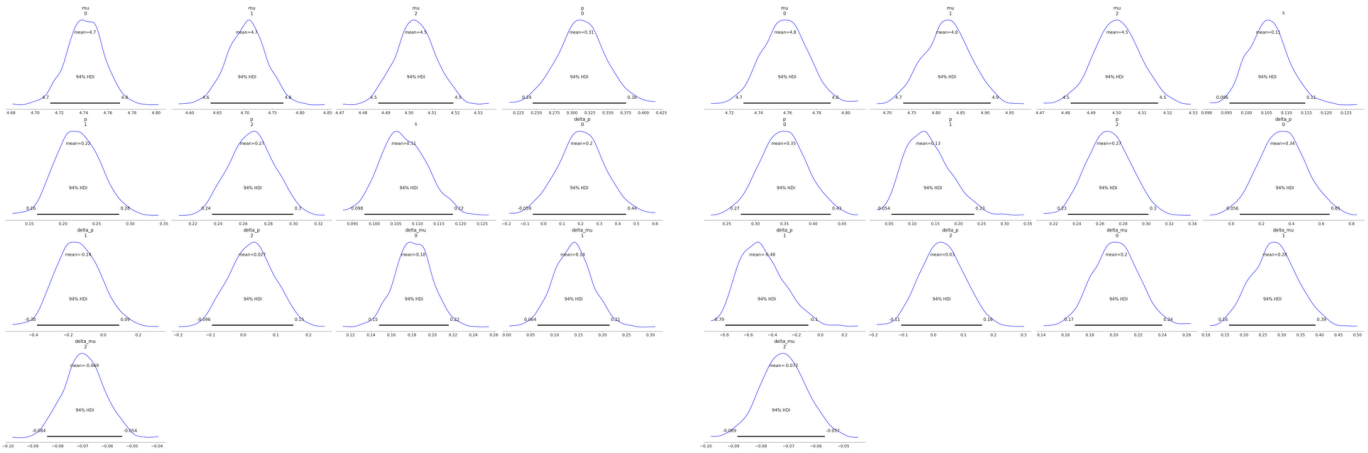


You can see from looking at the Delta-p and Delta-mu that the range for the trace is much larger (more uncertainty) and the peak of the curves happen in different points (different uplift results).

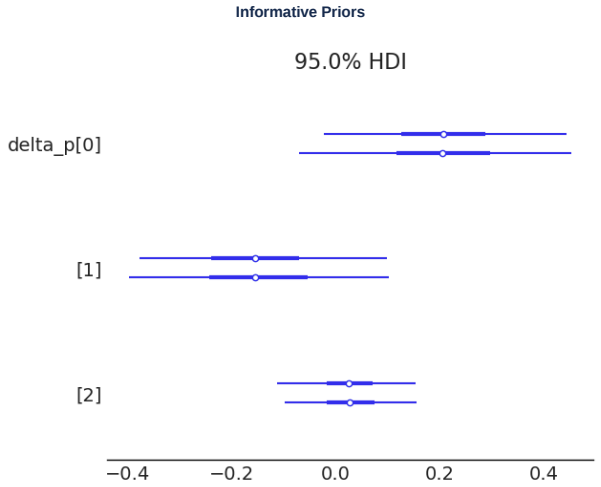
It also affects the posterior distributions:

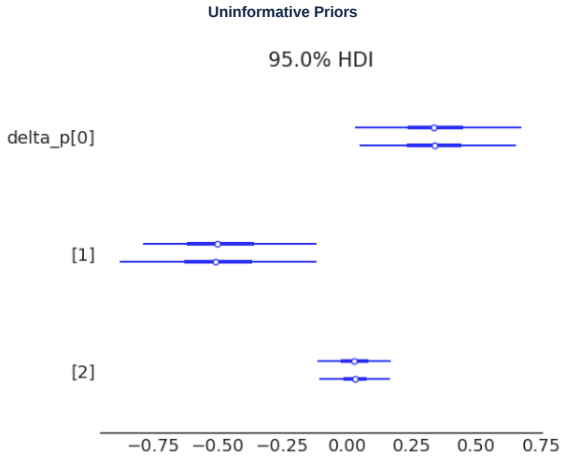
Informative Priors

Uninformative Priors



This also influences the HDI distribution and the range of values. We can see that the impact of Variations is smaller and the range of values is reduced. **We can also see that the results become statistically significant with Uninformative priors, but they are actually not when using Informative Prior.**





Which version of the test (A or B) is better? To answer this question, we approximate a distribution D which is the difference between the variations 1 (P). The probability that the variation is better is simply , or the integral of from 0 B PA P(D) 1 > 0 D1 to Infinity. Growth Book exposes this value as the “Chance to Beat Control.”

How much better is it? To answer this question, we approximate a second distribution D , which is the log of relative uplift.

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

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

$$R_B = \xi_A + \xi_B - E[P_B]$$

Gaussian quadratures (GQ): Gaussian quadrature is a numerical method used to approximate the definite integral of a function, particularly when the function is difficult to integrate analytically. It is especially effective for functions that can be well-approximated by polynomials.

Expected Loss: Bayesian Risk, also known as Expected Loss, is a concept used in Bayesian decision theory to quantify the expected value of the loss resulting from a decision or action, considering uncertainty and variability in the outcome. It integrates both the probability of various outcomes and the loss associated with each outcome.

Loss Function: A loss function quantifies the cost associated with a decision. The choice of loss function depends on the specific context and can significantly affect the decision-making process. Common examples include the squared error loss, absolute error loss, and zero-one loss.

Risk Calculation: Bayesian Risk (Expected Loss) for a decision is calculated by taking the weighted average of the losses associated with each possible outcome, where the weights are the probabilities of these outcomes.

Example:

Imagine a medical diagnosis scenario where a doctor must decide whether or not to administer a particular treatment based on test results. The loss function could represent the consequences of incorrect decisions, such as the cost of unnecessary treatment (if the treatment is given when not needed) or the cost of failing to treat a condition (if the treatment is not given when needed). Bayesian Risk here would be the expected cost considering the probabilities of various health states and the accuracy of the test results.

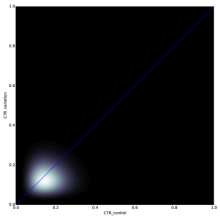
Joint Posterior

Suppose we have two page variants - say A and B. Suppose we have run an experiment, displaying variant A and B to nA and nB users. At this point we can compute a posterior for each variation - PA(λA) and B(λB).

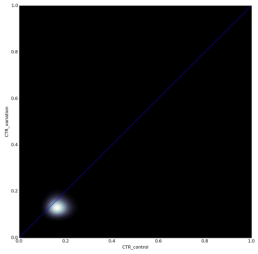
The joint posterior of A and B together is:

$$P(\lambda_A, \lambda_B) = P_A(\lambda_A) P_B(\lambda_B)$$

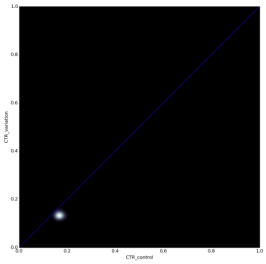
The joint posterior can be used to calculate various quantities of interest. The major quantities we are interested in are **loss functions** - functions which measure what sort of a mistake we will make assuming we choose a variant and stop the test.



Joint posterior near the start of the test. Points in the grey and white region represent highly likely values of (λA, λB), while dark regions represent areas of low probability for same. The blue line plots λA = λB.



Joint posterior near the middle of the test. The posterior is narrowing.



Joint posterior at the end of the test. The posterior has narrowed to the point where it lives almost entirely on one side of the blue line, and we can conclude that control is superior to the variation.

Chance to Beat Control

Suppose we have some evidence, and then we choose to display variant A. What iss the probability we made a mistake?

The Loss Function

The loss function corrects the error function in an important way. It treats small errors as less bad than big ones.

The loss function is the amount of uplift that one can expect to be lost by choosing a given variant, given particular values of λA and λB:

$$\mathcal{L}(\lambda_A, \lambda_B, A) = \max(\lambda_B - \lambda_A, 0)$$

As an example, suppose we choose to display variant A. Suppose the conversion rate for A is known to be λA = 0.1 and the conversion rate for B is λB = 0.15. Then the loss max(0.15 − 0.1, 0) = 0.05. In contrast, if we chose B, the loss would be max(0.1 − 0.15, 0) = 0.0.

The expected loss given a joint posterior is the expected value of the loss function:

$$E[\mathcal{L}](?) = \int_0^1 \int_0^1 \mathcal{L}(\lambda_A, \lambda_B, ?) P(\lambda_A, \lambda_B) d\lambda_B d\lambda_A$$

With ? meaning either A or B, depending on what you want to find out.

Running the test

The basic idea is to choose a desired error tolerance, which we denote by ε, a threshold of loss which is considered acceptable. Then the A/B test is run until the expected loss is below this specified tolerance.

The interpretation of ε is as follows - ε is a percentage. It represents how much lift one would expect to lose by making a particular choice given that the choice is wrong. It should be set to a number so low that one does not care if an error is made by this amount.

For example, suppose we are testing two button colors, and we are interested in measuring a lift of 10%. Conversely, if the lift we get from this test changes negatively by 0.2% or less, this change is so small that we don't care. In that case, we can choose ε = 0.002.

In terms of **sample size**, you can use the ROC-AUC curve to determine how precise you model is. The more data, the more precise it is.

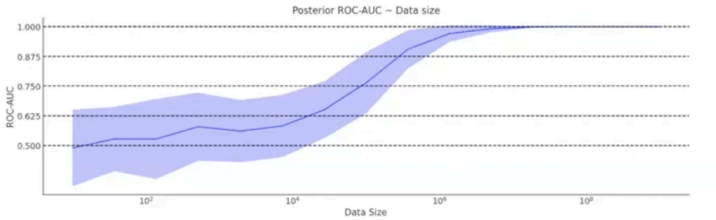


Figure: ROC-AUC increases as you get more data