Bayesian models for A/B experiments

Setup

We have conversion histograms for treatment and control groups, namely, for each group, we have a number of users who made a given observed number of conversions. The goal is to build Bayesian models using 1) aggregated treatment and control group data, and 2) histogram-level data.

0.1 Aggregated two-step Bayesian model

The simplest model uses aggregated level data:

- n_T and n_C - treatment and control group sizes
- c_T and c_C - number of converters (users who made a purchase) in treatment and control groups
- x_T and x_C number of conversions (purchases) in treatment and control groups

The model estimates p_T and p_C are the probabilities of conversion (becoming a converter), and λ_T and λ_C are the number of conversions per converter.

Uninformative priors are:

$$p_T \sim Beta(1,1),\ p_C \sim Beta(1,1)$$

$$\lambda_T \sim Gamma(1,1),\ \lambda_C \sim Gamma(1,1)$$

Note: for large group sizes ($> 100k$), distribution parameters for Priors do not change the resulting sample distribution.

Given the priors, the estimating procedure is:

Step 1: given n_T, n_C and observable number of converters, we model p_T and p_C as:

$$c_T \sim Binomial(n_T, p_T),\ c_C \sim Binomial(n_C, p_C)$$

Step 2: given number of converters c_T, c_C and number of conversions x_T, x_C, we model:

$$x_T \sim Poisson(c_T \cdot \lambda_T),\ x_C \sim Poisson(c_C \cdot \lambda_C)$$

The sampler (NUTS) generates a full posterior distribution for each of the aforementioned parameters. Thus, we get for the expected conversion per user:

$$E_T = p_T \cdot \lambda_T,\ E_C = p_C \cdot \lambda_C \ \Rightarrow\ Lift = E_T - E_C$$

Why use this model?

- Fast and simple
- interpretable

!But strong Poisson distribution assumption (mean = variance), which is never true. On practice works badly with highly overdispersion and underdispersed. Regular recommendation is to use the Zero-Inflated Negative Binomial distribution - a very unstable model with unpredictable results.

0.2 Dirichlet-Multinomial model

From the histogram, we observe:

- vector of possible purchase counts $\mathbf{k} = [0, 1, 2, ..., K_{max}]$
- histogram of counts $\mathbf{c} = [c_0, c_1, ..., c_{K_{max}}]$ with c_k being a number of users who made exactly k purchases
- $n = \sum_{k=0}^{K_{max}} c_k$ the total number of users in the group
- $\mathbf{e} = \frac{c}{n} = [e_0, e_1, ..., e_{K_{max}}]$ empirical probability vector

The model estimates a concentration parameter c, needed for model stability and for reproducing accurate results, and the vector of true probabilities $\mathbf{p} = [p_0, p_1, \ldots, p_{K_{\max}}]$, where p_k is the true probability that a random user makes exactly k purchases.

Priors are:

$$\log c \sim \mathcal{N}(1, 1)$$

being the concentration parameter, $c >= 0$.

$$\mathbf{p} = Dirichlet(\alpha), \text{ where } \alpha = c \cdot \mathbf{e} + \varepsilon,$$

with $\varepsilon = 10^{-3}$ being a small constant for numerical stability.

Given the priors, the $\mathbf{c}$ vector is modeled as a single draw from a Multinomial distribution with n total trials and the probability vector $\mathbf{p}$:

$$\mathbf{c} \sim Multinomial(n, \mathbf{p})$$

Expected rate per user is calculated by taking the dot product of the true probability vector $\mathbf{p}$ and the vector of purchase counts $\mathbf{k}$:

$$Rate = \sum_{k=0}^{K_{max}} k \cdot p_k$$

and the MCMC process generates a full posterior distribution for this Rate, which can then be used to compare the treatment and control groups (e.g., by calculating $\text{Rate}_T - \text{Rate}_C$).

Why to use this model?

- very robust and flexible without strict distribution assumptions
- very granular analysis

!But slow and not very interpretable.

0.3 Dirichlet-Poisson Mixture model

From the histogram, we observe for converters:

- N being the number of bins
- $\mathbf{k} = [k_1, k_2, ...k_N]$ modeled vector of conversions (purchases)
- $\mathbf{f} = [f_1, f_2, \ldots, f_N]$: The vector of frequencies, where f_i is the number of users who made exactly k_i purchases
- The overall conversion rate p_{conv} that goes as a fixed constant

- K - fixed number of latent subgroups (**crucial parameter for the model**)

The goal is to estimate 1) mixture weights $\mathbf{w} = [w_1, \ldots, w_K]$, a vector of probabilities describing the chance a random converter belongs to subgroup j (with $w_j \geq 0$ and $\sum_{j=1}^{K} w_j = 1$), and 2) component means $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]$, a vector where λ_j is the average purchase frequency for a user in subgroup j.

Uninformative priors are:

$$\mathbf{w} \sim \text{Dirichlet}(\mathbf{1}_K)$$

$$\text{For } j \in \{1, \ldots, K\} : \quad \lambda_j \sim \text{Gamma}(1.0, 1.0)$$

Such a prior distribution for $\mathbf{w}$ means that, a priori, we believe all K subgroups are equally likely to exist.

Given the priors, the procedure is:

The probability of k_i purchases, given they are from subgroup j, is given by the Poisson PMF:

$$P(k_i \mid \lambda_j) = \text{Poisson}(k_i \mid \mu = \lambda_j) = \frac{\lambda_j^{k_i} e^{-\lambda_j}}{k_i!}$$

The total probability for observing k_i is the sum of all K possibilities, weighted by the mixture weights w_j:

$$P(k_i \mid \mathbf{w}, \boldsymbol{\lambda}) = \sum_{j=1}^{K} w_j \cdot P(k_i \mid \lambda_j)$$

The **total log-likelihood** of the model is the sum of the log-likelihoods for all observed data points. Since the data is binned, the total log-likelihood is:

$$\log(\mathcal{L}) = \sum_{i=1}^{N} f_i \cdot \log\left(P(k_i \mid \mathbf{w}, \boldsymbol{\lambda})\right) = \log(\mathcal{L}) = \sum_{i=1}^{N} f_i \cdot \log\left(\sum_{j=1}^{K} w_j \cdot \frac{\lambda_j^{k_i} e^{-\lambda_j}}{k_i!}\right).$$

Thus, the model then finds the posterior distributions for $\mathbf{w}$ and $\boldsymbol{\lambda}$ that best explain the observed frequencies $\mathbf{f}$.

Expected purchases per user is thus calculated as

$$E[\lambda_{\text{conv}}] = \sum_{j=1}^{K} w_j \cdot \lambda_j,$$

with the population conversion rate being

$$\text{Rate}_{\text{pop}} = p_{\text{conv}} \cdot E[\lambda_{\text{conv}}]$$

with fixed observed p_{conv}. This model also provides an implied probability of being a converter:

$$P(\text{converter}) = 1 - \sum_{j=1}^{K} w_j e^{-\lambda_j}$$

which might be compared to p_{conv} to verify the model's fit.

Note: K parameter is unknown, choosing small K will underfit the data, large K can lead to overfitting and slow sampling. In practice, run the model multiple times with different K values and compare them using

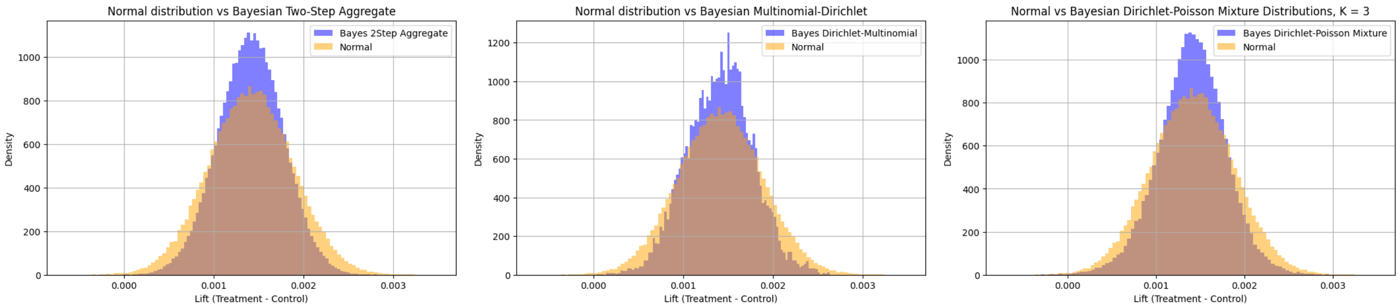
WAIC or LOO.

Why to use this model?

- handles overdispersion well
- can capture complex, multi-modal shapes
- relatively fast and interpretable, faster than Dirichlet-Multinomial for large histograms

Models comparison

In this section, I compare models' performances with a particular histogram example. To illustrate the models' performances, I plot the resulting sample for absolute lift. For the baseline, I use a normal distribution with parameters from a given histogram to mimic the frequentist approach. Since the group sizes are large, we may use CLT and expect the absolute lift distribution to be normal.



Samples from all three models have lower variance compared to the normal distribution. Means are the same with a slight shift to the right for the Dirichlet-Multinomial model. This observation demonstrates how Bayesian models can not only add flexibility to the interpretation but also shrink confidence intervals in the case of overdispersion.