AB Testing Using A Bayesian Decision Making Framework

Demetri Pananos

Introduction

This analysis intends to analyze the AB test provided from Voices.com. A decision must be made to either implement a change to the response page or continue with status quo. I intend to analyze this AB test by examining the cost of making a wrong decision. I use a heirarchical Bayesian logistic regression to model the hire rate for both variants, and then Bayesian decision analysis to compute the expected cost a wrong decision using the accompanying job data. Techincal details are included in an appendix. Considerations for extentions to the analysis are discussed.

Methods

The data show that members post multiple jobs during the experiment, violating i.i.d. assumptions required for z-test of proportions or similar tests. To combat this, I use a multilevel Bayesian logistic regression to estimate the population effects of the variant, accounting for the multiple observations of some members.

The brief indicates that the hire rate for status quo is between 60% and 70%. Because I have chosen a Bayesian framework, this information can be directly passed to the model to improve precision of downstream estimates. For more on model details, prior distributions, and model checking, please see the appendix.

The model is capable of producing estimates for hire rate for both variants. Examining estimates of hire rate in isolation doesn't tell the whole story. If the new variant's effect on the hire rate is highly uncertain, then there is an appreciable risk that Voices.com could implement a variant to the response page which actually hurts hire rate. To understand these risks more thoroughly, I use Bayesian decision making to estimate the expected decrease in hire rate when a wrong decision is made.

Results

Estimated Hire Rates

A total of 1399 out of 2129 jobs were labeled hired during the experiment. The table below shows total hired jobs stratified by variant.

Table 1: Summary of hired jobs in experiment.

	A	B	Together
Hired	722	677	1399
Total	1114	1015	2129

Shown in table 2 are the expected hire rates from the model and accompanying 95% credible intervals [1]. The model estimates that variant B has a superior hire rate as compared to variant A, however the expected hire rate is not markedly larger than version

[1] Credible intervals are the way a lot of people want to interpret confidence intervals. They are regions of parameter space where there is 95% probability that the effect lives. It would be completely correct to say "with 95% probability, the hire rate for version A is between ….".

A. The uncertainty in the difference in hire rates covers negative differences, meaning that there is a chance that B leads to lower hire rate than A in practice.

Table 2: Esimtated hire rates from the model under variants A and B.

	Estimate	Uncertainty
B-A	3%	-3% - 9%
A	68%	64% - 72%
B	71%	67% - 75%

Minimizing Expected Loss

Implementing a variant which may potentially decrease hire rate is a risk Voices.com should seek to minimize. Using the model from the previous section, we can estimate the expected amount by which the hire rate would decrease if we implemented the wrong variant. The results of this analysis are shown in table 3. In summary, if Voices.com kept with variant A but variant B truly was better, Voices.com would expect to lose out on a possible 3.06% increase to hire rate. Compare this to the other scenario, in which implementing variant B when variant A truly was better, Voices.com would experience a decrease in hire rate of 0.28%. In summation, if Voices.com wanted to ensure their risk of loss was lowest, implementing variant B is the best option.

Table 3: Expected decrease in hire rate under different scenarios. If the superior variant is implemented, the loss is 0. If the inferior variant is implemented, loss is non-zero. For example, if A is implemented but B is truly superior, the hire rate would be 3.06% lower than what it could have been. If B is implemented but A is truly superior, hire rate would only be 0.28% lower than what it could have been.

	A Better	B Better
A Implemented	0.00%	3.06%
B Impemented	0.28%	0.00%

"Cash Rules Everything Around Me" - Wu Tang Clan: Loss in Terms of Dollars

Variant B is estimated to result in smallest loss. This loss can be translated into dollars by using the job dataset. Of all jobs in that dataset, Voices.com is expected to earn $0.2 \times$ hire rate $\times$ sum of average bid dollars. From the model, I estimate that had Voices.com implemented variant B instead of variant A for all jobs in the job dataset assuming variant A is truly superior, then Voices.com have lost out on $8,372.90 worth of revenue. Compare this to the loss of $89,853.89 worth of revenue when sticking with variant A assuming variant B is truly better. This underscores how costly a wrong decision can be.

Table 4: Expected revenue loss under different scenarios.

	A Better	B Better
A Implemented	$0.00	$89,853.89
B Impemented	$8,372.90	$0.00

Answers to Questions From The Leadership

1. Will an update to the Voices.com's response page increase hire rate?

From the model, I estimate there is a 82% chance that implementing the update will increase hire rate. The expected change in hire rate would be an increase of 3% from 68% to 71%.

2. Are there any additional findings we can get from our job data?

Not only do we expect variant B to lead to superior hire rate, but implementing variant B over A also decreases expected loss.

I investigated expected revenue loss using the job data. Had variant B been applied to all jobs in the job data when variant A was truly better, Voices.com would have lost $8,372.90. Had variant A been applied to all jobs in the job data when variant B was truly better, Voices.com would have lost $89,853.89. The cost from choosing the wrong variant is smallest when variant B is chosen.

3. What Should We Do?

Assuming Voices.com seeks to minimize losses in addition to increasing hire rate, Voices.com should implement variant B. However, looping in the business and discussing risk tolerance is an important step in making the decision which I am not able to do presently.

Conclusions and Considerations

Variant B is estimated to have a superior hire rate as compated to status quo. Additionally, implementing variant B when variant A is truly superior leads to a smaller loss than implementing variant A when variant B is truly superior. I used the accompanying job data to demonstrate that implementing variant A could possibly be a $90,000 dollar mistake, where as implementing variant B would potentially be ten times less costly.

A more thorough approach would be to model the hired price as a function of the job details (e.g. quote min, quote max, rating, etc) and then estimate expected revenue loss for both variants integrating over all uncertainty. This would ecplicitly model all uncertainty in the decsion process and give the most faithful estimates of loss. Had I more time, this is the approach I would have taken.

Appendix

I'm including this appendix for posterity. I would not expect business stakeholders to read this, but were I to actually work for Voices.com I would inlclude such a section for reproducibility and accountability for and to my fellow analysts.

Model

Some members are observed several times. This allows for each members hire rate to be estimated, and then the population level hire rate to be estimated. The model is then

$$\text{log}\left(\frac{p_i}{1-p_i}\right) = \beta_{0,i} + \beta_1 x$$

Here, p_i is the hire rate for member i, $\beta_{0,i}$ is baseline hire rate on the log odds scale for member i, and β_1 is the log-odds ratio for the variant over status quo. The assumption here is that each member's hire rate is normally distributed on the log odds scale with mean equal to population baseline hire rate and some variance which requires estimation. Mathematically,

$$\beta_{0,i} \sim \mathcal{N}(\beta_0, \sigma^2)$$

The brief indicates that the baseline hire rate is between 60% and 70%. I interpret this as a 95% credible region, and thus use a prior for the baseline hire rate in which 95% of the probability mass can be found between 0.6 and 0.7. On the log odds scale, this corresponds to the following prior

$$\beta_0 \sim \mathcal{N}(0.61, 0.12)$$

In my own experience, AB tests do not result in changes to baseline rate larger than 10%. A prior on the effect of the variant should be centered at 0 (i.e. null effect) and should not go too far beyond 2 standard deviations from the mean. In the absence of any other information on the variant, a standard normal prior should suffice

$$\beta_1 \sim \mathcal{N}(0, 1)$$

Loss Function

Let a and b be the true underlying hire rates under variants A and B respectively. The expected loss would be

$$E[L](x) = \int_A \int_B L(a, b, x) f(a, b) da\, db$$

Here, f is the joint posterior density of the hire rates, and L is the loss function for implementing variant x

$$L(a, b, x) = \begin{cases} \text{max}(b - a, 0) & x = a \\ \text{max}(a - b, 0) & x = b \end{cases}$$

Given we implement variant x, L tells us how much the hire rate would decrease provided we implemented the worse variant.

Model Checking

To check my model, I draw from the posterior and determine how many jobs the model predicts to be hired. If the number of observed hired jobs is similar to the number of predicted hired jobs, this is an indication the model fits well.

Show below is a histogram of predicted hired jobs. We see that the observed hired jobs (red line) is very close to the mean of the posterior distribution of predicted hired jobs. Thus, we conclude our model fits appropirately. That the mean is lower than the observed is an expected feature since the multilevel nature of the model will pool estimates towards the population level hire rate.