Bayesian A/B Testing for Measuring User Engagement in a Mobile App

Arunav Saikia
M.S. Data Science
Indiana University - Bloomington
arsaikia@iu.edu

Abstract

A/B Testing is a popular framework for judging the effectiveness of new features or changes in internet based companies. In simple terms, one group (control) is shown an existing version of a feature and the second group (treatment) is shown the new version. The stakeholders then want to gauge the effect the proposed change has on one or several KPIs (key performance indicators) to take informed business decisions. Often statisticians and analysts use a frequentist approach to answer questions about the KPI(s) and the experiment, using p-values, statistical significance and confidence intervals which fail to answer many direct questions the stakeholders might have. The goal of this project is to leverage Bayesian statistics to answer specific question(s) about the KPI(s) and the experiment in general, using our prior beliefs and evidence from the experiment.

1 Summary

The mobile team for a reading app introduced a new UI design with the goal of increasing user engagement on the app. The team wanted to evaluate the change and understand the impact of the UI change better by running an A/B test. The stakeholders were very keen about the new UI change and had specific questions about the impact of the proposed change eg. - how much has the engagement improved as a result of the UI change, what is the probability that the engagement for the treatment group is higher than the control group etc.

To measure the effectiveness of the change on user engagement we came up with two metrics - 1. average active time per user, defined as the ratio of the total minutes spent on the app by a user to the total number of days the user was active and 2. average active time per day, defined as the ratio of total time spent on the app by all users who were active on a single day to the number of users who were active. While the first metric is at the granularity of every user, the second is at the granularity of each day. The two metrics were designed to capture the effectiveness of the change both in regards to the individual users and for every single day. Since the stakeholders were interested in understanding the impact of the UI change on user engagement, we defined these two metrics which will act as indicators for user engagement. We will use the raw data to derive the metrics and use updated posterior estimates to comment on the efficacy of the UI change. The key questions we want to answer using the two metrics are -

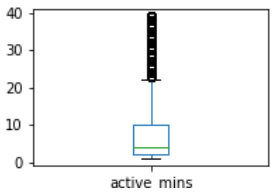

Figure 1: Box plot for active minutes from the raw data after outlier treatment

- What is the certainty that the new UI design worked and user engagement for the treatment group is better than the control group?
- How much has the user engagement for treatment group improved w.r.t the control group as a result of the UI change?
- Has the user engagement changed over the last few months as a result of the UI change. If so by how much?

2 Data

The raw data consists of minutes per day spent on the app by a sample of users every day for 180 days before and 150 days after the UI change. The

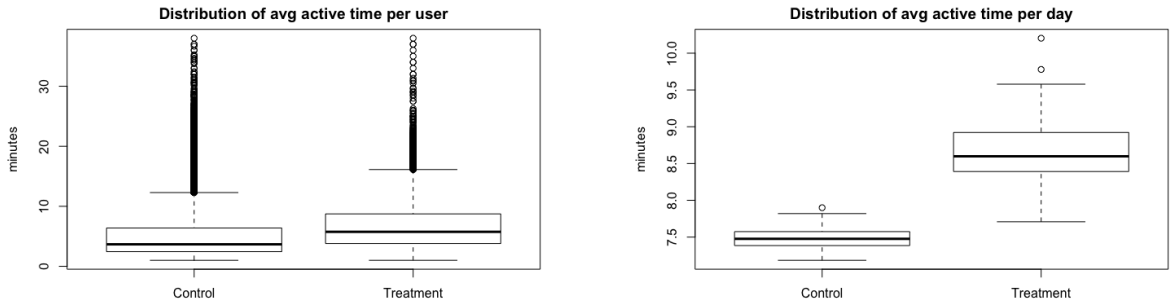

Figure 2: Distribution of the two metrics for the treatment and control groups

uid	date	treatment	active mins
10001	08/12/2020	1	5
10002	08/12/2020	0	6
10003	08/12/2020	0	2
10001	08/13/2020	1	6
10002	08/13/2020	0	2
10004	08/13/2020	0	10

Table 1: A sample of raw data

data has an identifier if a user was in the control or treatment group and other demographic information about the user like gender and user-type (eg. reader, non-reader, contributor etc). Approximately 37k users are in the control group and 9k in the treatment group. Table 1 shows a snapshot of the raw data. Figure 1 shows the data is right skewed with 75% activities on the app being less than 10 minutes long. We segment the data for our control and treatment group and use the raw data to generate our metrics. Distribution of the two metrics for both the groups can be seen in Figure 2.

3 Bayesian Analysis

3.1 Metric 1: Average active time per user

We assume our metric is generated from a normal distribution with mean μ and variance σ^2 having the following independent priors -

$$
\begin{aligned}
Y \mid \mu, \sigma^2 &\sim \mathcal{N}(\mu, \sigma^2) \\
\mu &\sim \mathcal{N}(\mu_0, \tau_0^2) \\
\sigma^2 &\sim \mathcal{IG}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})
\end{aligned}
$$

The full conditional distribution of the posteriors for this semi-conjugate model is given as follows -

$$\mu \mid \sigma^2, Y \sim \mathcal{N}(\frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}})$$

$$\sigma^2 \mid \mu, Y \sim \mathcal{IG}(\frac{n + \nu_0}{2}, \frac{\sum_{i=1}^{n}(y_i - \mu)^2 + \nu_0 \sigma_0^2}{2})$$

To determine our priors we performed an EDA on the raw data for 180 days before the experiment. By looking at the data for this period we saw that the empirical mean for the average active minutes per user was around 5 minutes with a standard dev of 4 minutes. Based on this we set the following priors $\mu_0 = 5$, $\tau_0^2 = 1$, $\sigma_0^2 = 16$. Since we did not want to give too much weight to our priors we set $\nu_0 = 1$. We ran the Gibbs sampler for 10000 iterations to get the posterior estimates for both the treatment and control group. Distribution of the posterior estimates for μ can be seen in Figure 3 (*left*).

Using the MCMC samples we determined that $Pr(\mu_{treatment} > \mu_{control} \mid Y_{treatment}, Y_{control}) = 99\%$ i.e. there is 99% probability that the average active time per user for the treatment group is higher than the control group. Using the distribution of the posterior differences we determined there is a 81% chance that average active time per user is higher for the treatment group than the control group by 1.5 minutes. On average, we observed the metric to improve by 1.28 times or 28% for the treatment group compared to the control group.

The above statistics showed that the UI change was successful, resulting in improved user engagement for the treatment group w.r.t the control. Next, we wanted to know if the user engagement changed

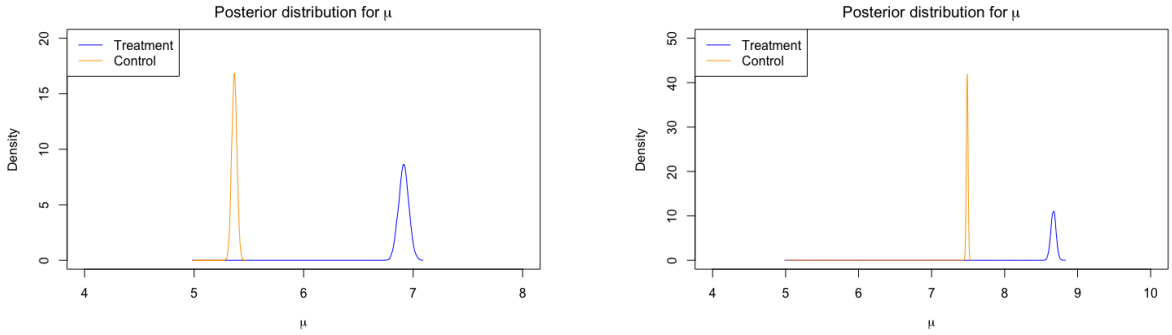

Figure 3: Marginal posterior distribution of the mean parameter for the two metrics, Metric 1 (left) and Metric 2 (right)

over the last few months as a result of the UI change. For this we measured the same metric for the users in the treatment group both before and after the change to check for improvement/decline. This is similar to a paired t-test in classical statistical analysis. For n users let,

Y_{i1} : average active time before UI change(i = 1,...n)
Y_{i2} : average active time after UI change(i = 1,...n)
We grouped these into a bi-variate user-level vector $Y_i = (Y_{i1}, Y_{i2})$. Therefore, we have Y_i is multivariate normal given as $Y_i \sim N_p(\mu, \Sigma)$ where $\mu = E[Y]$ a p-dimensional vector and $\Sigma = Cov(Y)$ is a $p \times p$ variance-covariance matrix. The semi conjugate priors for μ and Σ is

$$\mu \sim \mathcal{N}_p(\mu_0, \Lambda_0)$$
$$\Sigma \sim InvWishart_p(\nu_0, S_0^{-1})$$

The full conditional distribution of the posteriors is given as

$\mu \mid \phi, Y \sim Normal_p((\Lambda_0^{-1} + n\phi)^{-1}(n\phi\bar{Y} + \Lambda^{-1}\mu_0), (\Lambda^{-1} + n\phi)^{-1})$
$\phi \mid \mu, Y \sim Wishart_p(n + \nu_0, (S_0 + \sum_{i=1}^{n}(Y_i - \mu)(Y_i - \mu)^T)^{-1})$
where $\phi = \Sigma^{-1}$

We use the same prior for the means as before, $\mu_0 = (5, 5)^T$. Based on industry knowledge we know that the average active time per user on similar apps is likely between 3 and 7 minutes. We then set Λ_0 in a way there is small chance of it being outside this range

$$5 \pm 2\lambda_0 = (3, 7) \Rightarrow \lambda_0^2 = 1$$

This is the same reason we set $\tau_0^2 = 1$ in the previous analysis. Moreover, we can assume a strong prior correlation $\rho = 0.5$ and finally,

$$\Lambda_0 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

Since we have $\sigma_0^2 = 16$, with a prior correlation of 0.5 we have

$$S_0 = \begin{pmatrix} 16 & 8 \\ 8 & 16 \end{pmatrix}$$

To not commit too much to these choices, we'll set $\nu_0 = p + 2 = 4$. We ran the Gibbs sampler for 1000 iterations to get the posterior estimates.

Using the MCMC samples we determined that $Pr(\mu_1 < \mu_2 \mid Y_1, Y_2) = 99\%$ i.e. there is 99% probability that the average active time per user post UI change is higher than the before the change. Using the distribution of the posterior differences we determined there is a 96% chance that average active time per user is higher post the UI change than before by 2 minutes. On average, we observed the metric to improve by 1.43 times or 43% for post the change.

3.2 Metric 2: Average active time per day

Similar to Section 3.1 we can assume this metric to be univariate and normally distributed as well. To determine the priors we looked at the raw data for the same 180 day period before the experiment and observed that the empirical mean for the average active time per day was around 6.4 minutes with a standard dev of 0.16 minutes. Based on this we set the following priors $\mu_0 = 6.4$, $\tau_0^2 = 1$, $\sigma_0^2 = 0.0256$ and $\nu_0 = 1$. Distribution of the posterior estimates of μ for the treatment and control

groups, after running the Gibbs sampler for 10000 iterations can be seen in Figure 3 (*right*).

Using the MCMC samples we determined that $Pr(\mu_{treatment} > \mu_{control} \mid Y_{treatment}, Y_{control}) = 99\%$ i.e. there is 99% probability that the average active time per day for the treatment group is higher than the control group. On average we observe the metric to improve by 1.15 times or 15% for the treatment group compared to the control group.

4 Conclusion and Future Work

In this work we saw how we can apply Bayesian methodologies to deliver insights for AB testing. In contrast to classical frequentist approach of 2 sample t-tests, paired t-tests, p-values, and confidence intervals, a Bayesian approach enables us to answer direct questions about the KPI(s) of interest. We observed how we can use derived metrics like average active time per user and average active per day, to gauge effectiveness of a new UI design for improving user engagement. The key insights we observed are

- From both metrics we saw there is 99% probability that the user engagement for the treatment group is higher than the control group.
- There is a 81% chance that average active time per user is higher for the treatment group than the control group by 1.5 minutes.
- Average active time per day improved by 1.15 times or 15% for the treatment group compared to the control group.
- There is a 96% chance that average active time per user is higher post the new UI change than before by 2 minutes.

For this work we assumed that both the metrics are normally distributed. But from Figure 2 (*left*), we can see that metric 1 is skewed to the right. In future we can try performing a log transformation to the data and reduce the skewness. Moreover, we can use a gamma distribution as the sampling model with Metropolis Hastings to sample the posterior estimates. Also, since metric 2 is a day level metric, the samples are not independent and there exists autocorrelation which we have ignored. For this we can try working with Bayesian models specific to time series data.

We also assumed that the users are identical and exchangeable , but in reality, some users are avid readers on the app while some are contributors and others are non-readers. Consequently, the mean parameter for sampling model(s) might be different for the different cohorts and we can use hierarchical modeling to make posterior inference about the cohorts. For example stakeholders might be interested to know how has the UI change affected the engagement for each cohort (readers, contributors etc).

5 Source Code

Source code for all experiments and results can be found here `https://github.com/arunavsk/Bayesian-AB-Testing`.

6 Acknowledgements

The author wishes to thank Prof Daniel Manrique-Vallier for his instructions and support. The author also wishes to thank Miguel Pebes-Trujillo for his feedback and suggestions. This work was part of Bayesian Theory and Data Analysis (STAT-S 626) for Fall 2020 at Indiana University, Bloomington.

References

[1] Hoff, Peter D. A first course in Bayesian statistical methods. Vol. 580. New York: Springer, 2009.