Applying and Exploring Bayesian Hypothesis Testing for Large Scale Experimentation in Online Tutoring Systems

Vijaya Dommeti
Rivier University
Nashua, NH 03060, USA
vdommeti@rivier.edu

Douglas Selent
Rivier University
Nashua, NH 03060, USA
dselent@rivier.edu

ABSTRACT
This paper demonstrates the viability of using Bayesian hypothesis testing for statistical analysis of experiments run in online learning systems. An empirical Bayesian method for learning a genuine prior from past historical experiment data is applied to a dataset consisting of twenty-two randomized controlled A/B experiments collected from the ASSISTments online learning platform. We show that using only twenty-two experiments results in a learned genuine prior with poor confidence interval estimates, and that roughly 200 experiments are required for a reasonable estimate of the true probability of an experiment having differences between experiment groups. We also conducted a leave-one-experiment-out cross-validation experiment, where a genuine prior is learned from twenty-one of the randomized controlled experiments provided in the dataset and then used to evaluate the remaining experiment. From this experiment we show that Bayesian hypothesis testing performs similar to Frequentist hypothesis testing and both methods were in agreement.

Author Keywords
Bayesian Hypothesis Testing, ASSISTments, Large Scale Experimentation, Randomized Controlled Trials

INTRODUCTION
Within the last decade, online experimentation is becoming more popular in the area of learning sciences and educational research. Several intelligent tutoring systems and online learning platforms are now running randomized controlled experiments on a regular basis. A common reason for the growing popularity of these systems is grounded in research done by Bloom [2]. Bloom showed that a small student-teacher ratio has a large impact on the effect size for improving student learning (two standard deviations). This is now commonly known as "Bloom's 2 Sigma Problem". Many online tutoring systems attempt to achieve the same result as Bloom with a reduced cost, by having the computer tutor mimic the functionality of a one-to-one tutoring scenario with a teacher/tutor.

As a result of the increased popularity of online tutoring systems, there are a growing number of randomized controlled experiments being run in these systems. Online experimentation at large scale is an area initially studied by various web-facing companies such as Google, Microsoft, Facebook and others [7]. Many of these companies are moving toward using Bayesian hypothesis testing (as opposed to the traditional frequentist null hypothesis statistical testing) to analyze the results of their experiments.

There are several advantages to using Bayesian hypothesis testing over Frequentist hypothesis testing. One advantage of using Bayesian hypothesis testing is that it produces a probability of a null result in addition to probabilities for positive and negative effects [4]. This is contrary to frequentist statistics where we can only fail to reject the null hypothesis. As a result, a large number of scientific experiments are not reported or published because of the lack of ground-breaking findings. Most experiments do not result in conclusive results. Kohavi discusses in some cases up to 90% of experiments do not result in any changes in companies such as Google and Netflix [7]. This contributes to the "File Drawer" problem, where a large number of null results fail to be published due to publication bias, which favors positive results [10].

Using Bayesian statistics will also help alleviate the reproducibility problem, where the reported results of many research findings cannot be reproduced. Ioannidis has demonstrated this in the medical field by analyzing 49 clinical research studies reported in the top three medical journals. He has shown that out of the 49 studies, 45 reported that the intervention was effective. Out of these 45 interventions, 16% were contradicted by subsequent studies, 16% had found effects that were originally stronger than the effects of subsequent studies, 44% were replicated, and 24% remained unchallenged [8].

Ioannidis has stated that one of several reasons for the lack of reproducibility is because of a low prior probability of the research findings being true [9]. Standard statistical methods do not consider this piece of information, which results in a large number of false positive (Type-I error) results.

L@S 2017, April 20-21, 2017, Cambridge, MA, USA
 ISBN 978-1-4503-4450-0/17/04…$15.00
DOI: http://dx.doi.org/10.1145/3051457.3053992

In order to conduct Bayesian hypothesis testing, a *genuine prior* must be known or learned from data. This genuine prior represents the prior probability of an event being true. For example, a coin flip has a prior probability of 0.5 for the coin to land on heads. Those who are using Bayesian hypothesis testing for online controlled experiments are using historical data on thousands of previously run experiments to accurately learn a prior from those previously run experiments when such prior is not known.

In the context of this paper, the genuine prior will represent the prior probability of an A/B experiment having a difference in Mastery Speed between the control group and experiment group. Mastery Speed is the total number of problems a student attempted before reaching mastery and completing the assignment [12].

This paper demonstrates the viability and advantages of applying Bayesian hypothesis testing to analyze randomized controlled experiments being run in online tutoring systems. First we apply the objective Bayesian A/B testing framework described by Alex Deng to learn a Bayesian genuine prior from an unbiased collection of experiment data on twenty-two randomized controlled A/B tests obtained from the ASSISTments online learning platform [4, 11]. We then run a separate experiment, where we learn a genuine prior from twenty-one experiments and use that prior to perform Bayesian hypothesis testing on the twenty-second experiment in a leave-one-experiment-out cross-validation format. This is done for each of the twenty-two experiments, where one experiment is evaluated using the genuine prior learned from the other twenty-one experiments. We compare the results between the two statistical methods.

DATA DESCRIPTION

The dataset[1] we use in our experiments comes from 22 randomized controlled A/B tests run inside the ASSISTments online learning platform [11]. A total of 6,819 unique students participated in these experiments. All of the experiments were created by either internal or external researchers working with ASSISTments. There are two major characteristics of this dataset that make it ideal for conducting our experiments.

Firstly all of the 22 experiments are in a canonical format with the same dependent measure. All experiments were mathematic assignments with a control group and an experiment group. Students were randomly assigned into one of the two groups. Students continued to receive problems in the tutoring system until they had reached mastery of the content. Mastery occurs after a student has answered *n* problems correctly in a row, where *n* is typically set to three. The logarithm base ten of Mastery Speed is used as the dependent measure for the experiments in this dataset to reduce the effect of outliers on the mean.

Secondly these experiments represent an unbiased collection of experiments with positive, negative, and null results. Due to publication bias, it is hard to obtain such a dataset because positive results are reported more often than negative or null results. These two characteristics are important in order to learn a genuine prior, which requires an unbiased sample estimate of the population of experiment outcomes.

LEARNING A GENUINE PRIOR

We implemented the method described by Alex Deng to learn a genuine prior from the dataset [4]. This method works by first calculating effect sizes for each experiment. Since it is unknown whether these effect sizes were generated from experiments with a true difference between conditions (alternate hypothesis), expectation maximization is used to calculate the posterior odds of a given effect size belonging to the alternate hypothesis against the null hypothesis. The final learned posterior odds are then converted to a probability which is used as the genuine prior in Bayesian analysis. The details of this algorithm are described by Deng [4].

We calculate confidence intervals for the learned prior using the bootstrap method described in Basford et al [1]. We also simulate having a larger number of experiments to see how many experiments are required for reasonable confidence on the learned prior. To do this we vary the sample size of the bootstrap samples, where the sample size is equivalent to the number of experiments.

Figure 1 shows the confidence intervals for a varying number of experiments with 10,000 bootstrap samples for each sample size. The mean probability (for a given experiment to have an effect on Mastery Speed) learned back is roughly 0.39 for all numbers of experiments with varying confidence intervals. We acknowledge that the confidence intervals for the prior learned on twenty-two experiments are quite poor [0.001, 0.86]. The number of experiments required for a reasonable confidence interval is somewhat subjective; however it appears that there would need to be roughly 200 experiments of historical data to learn from.

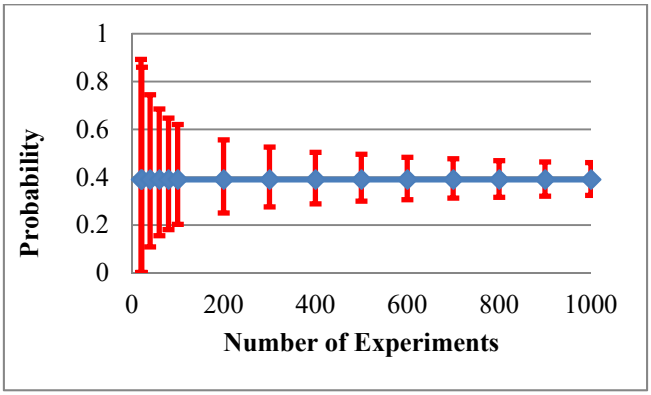

Figure 1. Confidence interval estimates on learned probability

[1]https://sites.google.com/site/las2016data/data

APPLYING BAYESIAN HYPOTHESIS TESTING

To apply Bayesian hypothesis testing we use leave-one-experiment-out cross-validation, using the methodology previously described to learn a prior with data from all of the experiments except one. We then use that learned prior to evaluate the last experiment. This repeats until all twenty-two experiments have been evaluated. We compare the results for both the Bayesian and Frequentist (T-Test) analysis methods.

Table 1 summarizes the results for all experiments. For the Frequentist results we report the *p*-value and the effect size (Cohen's D). In Bayesian statistics there are three probabilities that are reported. P(flat) is the probability of the result being zero, P(positive) is the probability of the result being positive, and P(negative) is the probability of the result being negative [4]. In our context P(flat) represents the probability of the treatment having no impact on student learning, P(positive) represents the probability of the treatment having a positive impact on student learning, and P(negative) represents the probability of the treatment having a negative impact on student learning.

In addition to the three probabilities, it is common to report the Bayes Factor. The Bayes factor is the ratio of the posterior odds of the alternate hypothesis being true to the prior odds of the alternate hypothesis being true [6]. For example, if the Bayes factor is ten, then this can be interpreted as the data is ten times more likely to occur under the alternate hypothesis. Kass & Raftery give general benchmarks on interpreting the Bayes Factor [6].

	Bayesian						Frequentist	
Experiment Number	Learned Prior	Learned Variance	Bayes Factor	P(positive)	*P*(flat)	P(negative)	*p*	Effect Size
1	0.445	0.199	0.348	0.218	0.782	0.000	0.812	0.017
2	0.362	0.215	4.383	0.713	0.287	0.000	0.016	0.186
3	0.367	0.192	414.230	0.996	0.004	0.000	0.000	0.464
4	0.336	0.200	44.597	0.958	0.042	0.000	0.001	0.290
5	0.359	0.199	0.303	0.000	0.855	0.145	0.606	-0.029
6	0.370	0.200	0.483	0.206	0.779	0.015	0.871	0.018
7	0.378	0.201	0.484	0.000	0.773	0.227	0.745	-0.034
8	0.384	0.201	0.439	0.000	0.785	0.215	0.768	-0.028
9	0.389	0.201	0.468	0.008	0.771	0.222	0.8475	-0.020
10	0.381	0.203	2.491	0.606	0.394	0.000	0.038	0.200
11	0.384	0.204	0.594	0.000	0.730	0.270	0.799	-0.085
12	0.387	0.204	0.528	0.000	0.750	0.250	0.663	-0.050
13	0.382	0.204	3.016	0.651	0.349	0.000	0.029	0.277
14	0.382	0.205	0.962	0.000	0.627	0.373	0.136	-0.120
15	0.382	0.205	0.844	0.000	0.657	0.343	0.2244	-0.135
16	0.382	0.206	0.948	0.370	0.630	0.000	0.261	0.207
17	0.383	0.206	0.762	0.321	0.679	0.000	0.245	0.118
18	0.386	0.206	0.367	0.004	0.813	0.183	0.8727	-0.013
19	0.386	0.206	0.746	0.320	0.680	0.000	0.146	0.088
20	0.390	0.205	0.225	0.082	0.874	0.044	0.985	0.001
21	0.392	0.205	0.471	0.000	0.767	0.233	0.654	-0.044
22	0.389	0.206	1.745	0.000	0.473	0.527	0.0653	-0.187

Table 1. The results comparing both Bayesian and Frequentist hypothesis testing. The Bayesian method used an initial prior learned from EM as well as an initial variance. All three probabilities are reported for Bayesian statistics as well as the Bayes factor. The *p* value and effect size (Cohen's D) are reported for the Frequentist statistics.

Overall the results reported in Table 1 are fairly standard results that would be expected from an unbiased collection of randomized controlled A/B experiments run in an online tutoring system. Both Bayesian and Frequentist methods report a null result for most of the experiments. Most effect sizes and Bayes Factors are small, which is expected in this area of research. Bayesian methods perform similar to Frequentist methods.

When comparing both the Bayesian and Frequentist methods we look to see if there are any differences and where those differences are. There are five experiments (2-4, 10, 13) with reported p-values < 0.05. Out of these five experiments only two (2, 3) had a P(positive) > 0.95 with the Bayesian method. Those two experiments were the only two experiments with either a P(positive) or P(negative) greater than 0.95. Without knowing the ground truth values it is not possible to say which method is more accurate. It is also worth pointing out that although Bayesian methods can accept the null hypothesis, none of the experiments had a P(flat) > 0.95. Fortunately there was only one experiment with a P(negative) of greater than 0.5, which was experiment 22.

There are several P(flat) probabilities between 0.7 and 0.9. Although these probabilities are not yet large enough to accept the null, they provide promise that with possibly a few more samples the probabilities would be large enough to accept the null hypothesis; Thus turning a null result by Frequentist methods into a conclusive result by using Bayesian methods.

FUTURE WORK

The work described in this paper applies Bayesian hypothesis testing on existing experiment data. Since it was shown that Bayesian hypothesis testing can be used, given enough historical experiments, it is future work to continue to implement this framework into online tutoring systems to better manage the randomized controlled experiments. It is expected that as the number of historical experiments increase, Bayesian hypothesis testing will be a superior alternative to traditional methods. Several Bayesian advantages can then be gained, such as continuous monitoring and optional stopping. There already exists a template to implement these methods described in [3]. It is future work to continue to integrate Bayesian hypothesis testing into online tutoring systems for better experiment methodology that could ultimately benefit thousands of students using these systems.

CONTRIBUTIONS AND CONCLUSIONS

This paper makes a first attempt at applying Bayesian hypothesis testing to randomized controlled A/B experiments run inside online tutoring systems. We show that the methods can be applied in this context and extended the existing methodology to do so. We report on how to calculate confidence intervals on the prior probability learned from using expectation maximization on the effect sizes of treatments in experiments. We also show how many experiments would be required to have reasonable confidence estimates on the genuine prior learned from historical experiment data. We show that using Bayesian hypothesis testing generates results consistent with the Frequentist methods in addition to having a more intuitive probability interpretation.

REFERENCES

1. Basford, K. E., Greenway, D. R., McLachlan, G. J., & Peel, D. (1997). Standard Errors of Fitted Component Means of Normal Mixtures. Computational Statistics, 12(1), 1-18.
2. Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational researcher, 13(6), 4-16.
3. Deng, A. Lu, J., Chen, S. (2016) Continuous Monitoring of A/B Tests Without Pain: Optional Stopping in Bayesian Testing (ArXiv ver.)
4. Deng, A. (2015, May). Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments. In Proceedings of the 24th International Conference on World Wide Web (pp. 923-928). ACM.
5. Efron, B. (2013). A 250-year argument: Belief, Behavior, and the Bootstrap. Bulletin of the American Mathematical Society, 50(1), 129-146.
6. Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773-795.
7. Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013, August). Online Controlled Experiments at Large Scale. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1168-1176). ACM.
8. Ioannidis, J. P. A. (2005). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. JAMA: the Journal of the American Medical Association 294 (2): 218–228. doi:10.1001/jama.294.2.218.
9. Ioannidis J. P.A. (2005). Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124.
10. Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. Psychological Bulletin, 86(3), 638.
11. Selent, D., Patikorn, T., & Heffernan, N. (2016, April). ASSISTments Dataset from Multiple Randomized Controlled Experiments. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale (pp. 181-184). ACM.
12. Xiong, X., Li, S., & Beck, J. E. (2013, May). Will You Get It Right Next Week: Predict Delayed Performance in Enhanced ITS Mastery Cycle. In FLAIRS Conference.