Evaluation of Early Stopping in A/B Testing

Shan Huang

May 18, 2017

This documentation shows the evaluation of different early stopping algorithms.

1 Problem

Given samples **x** from treatment group, samples **y** from control group, We are interested in whether there is a significant difference between the mean the two variants $\delta = \mu(y) - \mu(x)$. To save the cost of long-running experiments, we want to stop early if we are already certain that there is a significant result.

2 Significance Analysis

Given H_0 represents the null hypothesis of no difference, H_1 represents the alternative hypothesis meaning there is a difference, we can draw the conclusion of whether the result is statistically significant using either:

- **Confidence interval**: If 0 is outside confidence interval of δ, it is statistically significant. Vice versa.
- or **credible interval**: Credible interval is the Bayesian version of confidence interval. If 0 is outside credible interval of δ, it is statistically significant. Vice versa.
- or **Bayes factor**: Bayes factors higher than 3 can be interpreted as support for the alternative hypothesis (significant difference), whereas values smaller than 1/3 can be interpreted as support for the null hypothesis (significant no difference). Values between 1/3 and 3 are inconclusive.

The ability of each metric, i.e., types of significance it can detect, is shown in the following table.

	paradigm	significant H_1	significant H_0	no significant result
confidence interval	frequentist	✓		✓
credible interval	Bayesian	✓		✓
Bayes factor	Bayesian	✓	✓	✓

For simplicity, we draw a binary conclusion that either there is a significant difference or not. In other words, the first column means that there is a significant difference, combining the second and third column means that there is no significant difference.
It is worth noting that a **typical conclusion of Bayes factor** would be "*There is a significant difference corresponds to Cauchy prior and a threshold of Bayes factor=3*". This might be quite difficult to explain to non-tech users. While the **conclusion based on interval** can be more intuitive such as "*You can be 95% sure that the significant difference is not due to chance*".

3 Early Stopping Criteria

We can stop the experiment by either

- **Confidence interval**: If 0 is outside confidence interval of δ, stop. Calculate the significance level for each day based on group sequential method.
- or **credible interval**: If 0 is outside credible interval of δ, stop.
- or **Bayes factor**: If Bayes factor is higher than 3 or smaller than 1/3, stop.
- or **Bayes precision**: If width of credible interval is smaller than 0.08, stop.

4 Evaluation

4.1 Evaluate Significance Analysis

We are going to compare FP, TP, FN, TN for the three significance analysis algorithms described in section 2.

4.2 Evaluate Early Stopping Criteria

It is obvious that if we use the frequentist approach in early stopping, we should also use confidence interval to make conclusion of significance. However, things get a bit more complicated in the Bayesian case. When using Bayes factor to stop, we often find a conflicting result if we draw significance based on credible interval. Find below a table of combination of early stopping algorithms and significance analysis algorithms we will evaluate on.

Significant Based On	Early Stopping Based On			
	Confidence Interval	Credible Interval	Bayes factor	Bayes precision
Confidence Interval	✓			
Credible Interval		✓		
Bayes factor			✓	✓

For each combination shown in the table above, we will evaluate the following metrics:

- **False positive rate**: Percentage of wrongly stopped experiment. i.e., Early stop says there is a significant difference and stops the experiment, but a regular test will tell you not significant when the sample size is reached.
- **Run time reduced**: Average run time reduced for the correctly stopped experiment.
- **Bias**: Percentage of the difference of effect size (delta/std).
- **False positive run time reduced**: Average run time reduced for the wrongly stopped experiment.

5 Result

5.1 Simulation Data

We generate 2000 simulation tests based on Gaussian distributed KPIs. We calculate the minimal detectable effect size calculated from power analysis. 1000 tests should have a significant difference between control and treatment, and the other 1000 tests should have no significant difference.
We simulate the A/B testing for a period of 20 days, where the frequency of the visit from an entity is modelled by a Poisson distribution with visits on 3 days per entity in average. We run an analysis and evaluate whether to stop on each day.

5.2 Real Data

Use real data from previous A/B testings.

6 Conclusion

Use group sequential and confidence interval.