
Evaluation of Early Stopping in A/B Testing

Shan Huang

May 23, 2017

This documentation shows the evaluation of early stopping algorithms.

1 PROBLEM

Given samples \mathbf{x} from treatment group, samples \mathbf{y} from control group, we want to know whether there is a significant difference between the means $\delta = \mu(y) - \mu(x)$.

To save the cost of long-running experiments, we want to stop the test early if we are already certain that there is a statistically significant result.

2 SIGNIFICANCE DECISION

We can decide whether the difference is statistically significant using either:

- **Confidence interval:** If 0 is outside confidence interval of δ , it is statistically significant.
- or **credible interval:** Credible interval is the Bayesian version of confidence interval. If 0 is outside credible interval of δ , it is statistically significant.
- or **Bayes factor:** Theoretically, Bayes factors higher than 3 can be interpreted as support for the null hypothesis (significant no difference), whereas values smaller than 1/3 can be interpreted as support for the alternative hypothesis (significant difference). Values between 1/3 and 3 are inconclusive. **In our test, we will only use Bayes factor less than 1/3 as decision criteria.** See reasons below.

Given the null hypothesis H_0 representing no difference, the alternative hypothesis H_1 representing a difference of the means, 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 | ✓ | ✓ | ✓ |

Table 2.1: Comparison of Decision Ability

To be consistent with all methods, 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. Thus, we only use the part where Bayes factor less than 1/3 as decision criteria.

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:** Calculate the new significance level for each day based on alpha-spending function in group sequential method. Stop the test if 0 is outside confidence interval of δ .
- or **credible interval:** Stop the test if 0 is outside credible interval of δ .
- or **Bayes factor:** Stop the test if Bayes factor is smaller than 1/3 (significant difference).
- or **Bayes precision:** Stop the test if credible interval width is smaller than 0.08.

4 METRIC

4.1 EVALUATE SIGNIFICANCE DECISION

We are going to compare false positive(type I error), false negative(type II error), true positive, true negative rates for the three significance decision criteria 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 frequentist's approach (that is confidence interval) to decide significance. Moreover, when using Bayes factor to stop, we often find a conflicting result if we draw significance conclusion based on credible interval. Therefore we don't evaluate all combinations of early stopping algorithms and significance decision algorithms, find below a table of combination we will evaluate.

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

Table 4.1: Combinations of Early Stopping and Significance Decision To Evaluate

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

- **False positive rate:** Percentage of tests which are wrongly stopped. i.e., Early stop says there is a significant difference and stops the test, but the test to the end day (as if there is no early stopping) will tell you it is not significant.
- **Run time reduced(all):** Percentage of run time reduced for all tests.
- **Run time reduced(true positive):** Percentage of run time reduced for only the correctly stopped tests.
- **Bias:** Difference of effect size(delta) between stopping day and end day of test(as if there is no early stopping).

5 EVALUATION

The evaluation is conducted on both simulation data and real data. Code can be found [here](#).

5.1 SIMULATION DATA

We generate 2000 simulation tests based on Gaussian distributed KPIs. We calculate the minimal detectable effect size from power analysis. 1000 tests have a significant difference between control and treatment, and the other 1000 tests have no significant difference.

We model the frequency of visits of an entity by a Poisson distribution with 3 days per entity in average. We simulate the A/B testing for a period of 20 days, and decide whether to stop on each day.

First of all, let's see the performance of different **significance decision** algorithms.

| | FPR | TNR | FNR | TPR |
|----------------------------|------|-------|-------|-------|
| confidence interval | 4.6% | 95.4% | 0.9% | 99.1% |
| credible interval | 4.6% | 95.4% | 1.0% | 99.0% |
| Bayes factor | 0.2% | 99.8% | 18.9% | 81.1% |

Table 5.1: Performance of Significance Decision on Simulation Data

We can observe in Figure 5.1 that Bayes factor works very good on false positive rate, but the two interval based approaches works better on true positive rate. In general, the two interval

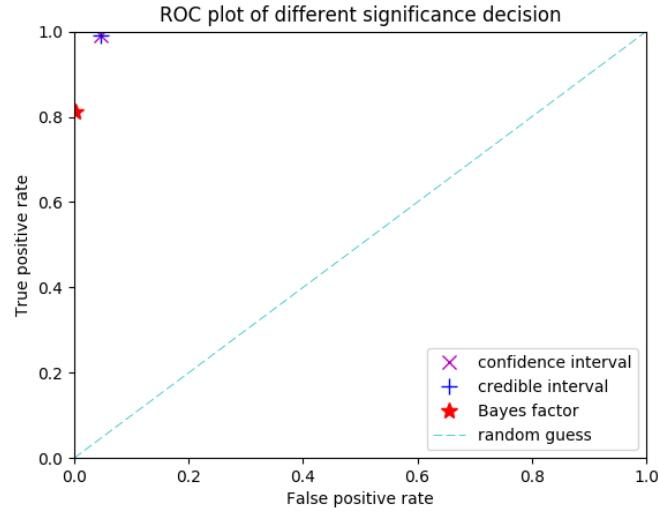


Figure 5.1: ROC plot on simulation data

based approaches are a little bit better because of larger distances to the line of random guess, while the error rate of all three methods are acceptable (FPR < 5% and FNR < 80%).

Next, let's look at the results of **early stopping** algorithms. To make notations compact, in the following table we use FPR for false positive rate, RTR(all) for run time reduced(all), RTR(TP) for run time reduced(true positive) and CI for confidence interval. The definition of each metric is described previously in Section 4.2.

| Significance | Early Stopping | Paradigm | FPR | RTR(all) | RTR(TP) | Bias |
|-------------------|-------------------|-------------|--------------|----------|---------|-------|
| CI | CI | frequentist | 3.3% | 1.60% | 16.74% | 0.00 |
| Credible Interval | Credible Interval | Bayesian | 22.6% | 17.56% | 42.28% | 0.00 |
| Bayes factor | Bayes factor | Bayesian | 0.7% | 0.57% | 42.5% | -0.02 |
| Bayes factor | Bayes precision | Bayesian | 0.2% | 78.02% | 78.02% | 0.00 |

Table 5.2: Performance of Early Stopping in A/A test on Simulation Data

| Significance | Early Stopping | Paradigm | FPR | RTR(all) | RTR(TP) | Bias |
|-------------------|-------------------|-------------|--------------|----------|---------|-------|
| CI | CI | frequentist | 1.7% | 51.58% | 51.74% | -0.01 |
| Credible Interval | Credible Interval | Bayesian | 0.7% | 78.53% | 78.78% | 0.00 |
| Bayes factor | Bayes factor | Bayesian | 7.8% | 46.44% | 52.96% | 0.00 |
| Bayes factor | Bayes precision | Bayesian | 67.8% | 77.95% | 77.76% | -0.02 |

Table 5.3: Performance of Early Stopping in A/B test on Simulation Data

5.2 REAL DATA

ToDo: Repeat the evaluation process on real data from previous A/B testings.

6 CONCLUSION

ToDo: Choose a significance decision criteria and an early stopping algorithm.