IEEE Conference

Sequential A/B testing

Yana Bondarenko
Department of Statistics and Probability Theory
Oles Honchar Dnipro National University
Dnipro, Ukraine
yana.bondarenko@pm.me

Abstract**—***In this paper sequential analysis to conversion rate optimization is considered. Sequential testing as the technique of decision making during A/B testing by sequentially gathering and processing the data is described. Sequential A/B testing results are presented descriptively and graphically.*

Keywords**—***A/B testing, composite hypothesis, conversion rate optimization, landing page, log-likelihood function, sequential analysis*

I. Introduction

Landing page optimization is the process of improving elements on a website to increase conversions. Landing page optimization is a subset of conversion rate optimization, and involves using methods such as A/B testing to improve the conversion goals of a given landing page. A/B testing is an experiment where two variants of landing page are shown to users at random, and statistical hypothesis testing is used to determine which variant performs better for a given conversion goal. For more details, see [1], [2], [3], [4] and [5].

In classical hypothesis testing the data collection is implemented without analysis and study of the data. Once all data are collected the statistical analysis is performed and conclusions are reached. Moving away from classical hypothesis testing to sequential analysis developed by A. Wald [6] was presented in papers [7], [8], [9], [10]. In sequential analysis every observation is analyzed immediately after being collected. The data are collected until that moment is then compared with preassigned threshold values, which integrate the knowledge obtained from the newly collected observation. This technique allows to reach conclusions during the data collection. Final conclusion can be reached with much fewer observations than in the case of classical hypothesis testing.

In this paper sequential analysis as the technique of decision making during A/B testing by sequentially gathering and processing the data is presented. M. Girshick's approach [11] to the composite hypothesis testing is examined and applied for A/B testing for the first time.

II. Sequential A/B Testing

A. Problem Formulation

Bernoulli trials with two possible outcomes (success and failure) are conducted in baseline and experimental groups of visitors. Probabilities of success p_1, p_2 are non-random unknown variables in baseline and experimental group, respectively. Hypothesis about probabilities of success is formulated as composite hypothesis $H : p_1 \le p_2$. Alternative hypothesis is also composite one $H' : p_1 > p_2$. Our goal is to accept or reject hypothesis H according to samples from Bernoulli distribution with parameters p_1, p_2 .

B. Sequential Analysis

Let p_1^0, p_2^0 be certain values of probabilities of success p_1, p_2 , respectively, where p_1^0, p_2^0 are chosen as $p_1^0 < p_2^0$. Let $H_0 : p_1 = p_1^0, p_2 = p_2^0$ be the hypothesis about the joint distribution $P(x_1, p_1^0)P(x_2, p_2^0)$ for the number of successes in one trial in baseline and experimental groups:

$$(p_1^0)^{x_1}(1-p_1^0)^{1-x_1}(p_2^0)^{x_2}(1-p_2^0)^{1-x_2},\ x_1 = 0,1;\ x_2 = 0,1 .$$

Let $H_1 : p_1 = p_2^0, p_2 = p_1^0$ be the hypothesis about the joint distribution $P(x_1, p_2^0)P(x_2, p_1^0)$ for the number of successes in one trial in baseline and experimental groups:

$$(p_2^0)^{x_1}(1-p_2^0)^{1-x_1}(p_1^0)^{x_2}(1-p_1^0)^{1-x_2},\ x_1 = 0,1;\ x_2 = 0,1 .$$

Sequential statistical technique applies to test simple hypothesis H_0 against simple alternative H_1 . Hypothesis H is accepted or rejected according to the outcome of the testing: hypothesis H_0 is accepted or rejected.

Two constants A, B must be chosen so that $0 < B < A$ and ratio of distributions on each stage of experiment should be evaluated for sequential hypothesis testing:

$$\frac{p_{1n}}{p_{0n}} = \frac{P(x_{11}, p_2^0)P(x_{21}, p_1^0)...P(x_{1n}, p_2^0)P(x_{2n}, p_1^0)}{P(x_{11}, p_1^0)P(x_{21}, p_2^0)...P(x_{1n}, p_1^0)P(x_{2n}, p_2^0)} ,$$

where x_{ki} is the i th observation for $x_k (k = 1, 2)$.

Let us suppose that pairs of observations $\left(x_{1i}, x_{2i}\right)$ are occurred, where each pair consists from one observation of number of successes x_1 in one trial in baseline group and one observation of number of successes x_2 in one trial in experimental group. Experiments continue until ratio p_{1n} / p_{0n} remains within interval $\left(B, A\right)$. Hypothesis H is accepted if $p_{1n} / p_{0n} \le B$ and rejected if $p_{1n} / p_{0n} \ge A$.

Ratio of distributions in testing simple hypothesis against simple alternative has the following form:

$$\frac{p_{1n}}{p_{0n}} = \left(\frac{p_2^0\left(1-p_1^0\right)}{p_1^0\left(1-p_2^0\right)} \right)^{\sum_{i=1}^{n}\left(x_{1i}-x_{2i}\right)} .$$

The log-likelihood function is equal to:

$$\ln\frac{p_{1n}}{p_{0n}} = \sum_{i=1}^{n}\left(x_{1i}-x_{2i}\right)\ln\left(\frac{p_2^0\left(1-p_1^0\right)}{p_1^0\left(1-p_2^0\right)} \right) .$$

M. Girshick [10] obtained that probability that sequential testing is completed with the acceptance of simple hypothesis H_0 depends only from value $\nu(p_1, p_2)$, this value can be interpreted as acceptable measure of deviation of p_1 from p_2 :

$$\nu(p_1, p_2) = \ln\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right) .$$

Let us suppose that we want to obtain testing process with following conditions: probability of rejecting the hypothesis H should not exceed a preassigned value α for whole area:

$$\nu(p_1, p_2) \le -d ,$$

and probability of the acceptance of the hypothesis H should not exceed a preassigned value β for whole area:

$$\nu(p_1, p_2) \ge d .$$

After that, we choose values p_1^0, p_2^0 in the way that

$$\nu(p_1^0, p_2^0) = -d .$$

The log-likelihood function takes the following form:

$$\ln\frac{p_{1n}}{p_{0n}} = d\sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right) .$$

Using

$$\frac{1}{d}\ln\frac{p_{1n}}{p_{0n}} = \sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right)$$

instead of the ratio p_{1n} / p_{0n} produces such testing scheme. We continue to simulate pairs of observations till inequalities are fulfilled:

$$\frac{\ln B}{d} < \sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right) < \frac{\ln A}{d} .$$

Hypothesis H is accepted if

$$\sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right) \le \frac{\ln B}{d} ,$$

and rejected if

$$\sum_{i=1}^{n}\left(x_{1i} - x_{2i}\right) \ge \frac{\ln A}{d} .$$

III. Classical A/B Testing

A. Problem formulation

Bernoulli trials with two possible outcomes (success and failure) are conducted in baseline and experimental groups of visitors. Probabilities of success p_1, p_2 are non-random unknown variables in baseline and experimental group, respectively. Hypothesis about probabilities of success is formulated as simple hypothesis $H: p_1 = p_2$. Alternative hypothesis is also simple one $H': p_1 - p_2 = \theta$. Our goal is to accept or reject hypothesis H according to samples from Bernoulli distribution with parameters p_1, p_2 .

B. Two Proportion Z-test

We used two-proportion *Z*-test for hypothesis testing. Frequencies of success $\hat{p}_1 = \mu_1/n_1, \hat{p}_2 = \mu_2/n_2$ are unbiased, consistent estimates for p_1, p_2 .

The weighted estimate of p_1, p_2 is defined by the formula:

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} ,$$

where n_1, n_2 are sample sizes for baseline and experimental groups, respectively. The null hypothesis $H: p_1 = p_2$ is rejected if Z statistic is greater than critical value $z_{1-\alpha/2}$ based on the standard normal distribution:

$$Z = \frac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \ge z_{1-\alpha/2} ,$$

or, that is the same, the null hypothesis $H: p_1 = p_2$ is rejected if difference $\hat{p}_1 - \hat{p}_2$ isn't contained in the confidence interval:

$$\left(-z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) .$$

C. Minimum sample size

If simple hypothesis $H: p_1 = p_2$ is true, then Z statistic has a standard normal distribution with large enough sample sizes n_i, n_j :

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\hat{p}_1 - \hat{p}_2}{SE} .$$

If alternative simple hypothesis $H': p_1 - p_2 = \theta$ is true, then Z_1 statistic has a standard normal distribution with large enough sample sizes n_1, n_2 :

$$Z_1 = \frac{\hat{p}_1 - \hat{p}_2 - \theta}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\hat{p}_1 - \hat{p}_2 - \theta}{SE} .$$

The power is the probability that test rejects the null hypothesis H when an alternative hypothesis H' is true:

$$P\{S \mid H'\} = P\left\{Z_1 \ge z_{1-\alpha/2} - \frac{\theta}{SE}\right\} + P\left\{Z_1 \le -z_{1-\alpha/2} - \frac{\theta}{SE}\right\} .$$

Let the power $P\{S \mid H'\}$ is equal to $1-\beta$. Then neglecting the second part, we have

$$P\left\{Z_1 \ge z_{1-\alpha/2} - \frac{\theta}{SE}\right\} \approx 1-\beta ,$$

$$P\left\{Z_1 < z_{1-\alpha/2} - \frac{\theta}{SE}\right\} \approx \beta .$$

By definition of the distribution function

$$N_{0;1}\left(z_{1-\alpha/2} - \frac{\theta}{SE}\right) \approx \beta ,$$

hence,

$$z_{1-\alpha/2} - \frac{\theta}{SE} \approx z_{\beta} ,$$

$$z_{1-\alpha/2} - z_{\beta} \approx \frac{\theta}{SE} ,$$

where z_{β} is β-quantile of standard normal distribution.

Let sample sizes n_1, n_2 be equal to n, then

$$z_{1-\alpha/2} - z_{\beta} \approx \frac{\theta}{\sqrt{\frac{2\hat{p}(1-\hat{p})}{n}}} .$$

From equality above follows that minimum sample size of the baseline group, required to prove that the probabilities of success in baseline and experimental groups of visitors are statistically different, is defined by:

$$n = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \hat{p}(1-\hat{p})}{\theta^2} ,$$

where α is the probability of making a Type I error, $1-\beta$ is the power of hypothesis testing, θ is expected improvement difference, $\hat{p}$ is the probability of success in baseline group.

IV. Sequential A/B Testing Results

A. Sequential A/B Testing Implementation

Landing page variant A is suggested viewing for the baseline group of visitors and landing page variant B is suggested viewing for the experimental group of visitors. We need to identify visitors during testing for clear experiment and suggest them the same landing page variant that they viewed earlier in case of repeated visits. The flow of visitors has been simulated. Each visitor can belong to the baseline group with probability $1/2$ and to the experimental group with probability $1/2$. Visitor behavior is simulated after identification of visitor belonging to group. Visitor behavior is determined with two outcomes: success – conversion action is done, failure – conversion action isn't done. If visitor belongs to the baseline group, success will happen with probability p_1 , and if visitor belongs to the experimental group, success will happen with probability p_2 . Composite hypothesis $H: p_1 \le p_2$ is formulated and sequential probability ratio test is used for hypothesis testing.

Let $p_1^0, p_2^0, \alpha, \beta$ be preassigned values. The constants A, B are equal to $A = (1-\beta)/\alpha$, $B = \beta/(1-\alpha)$. Experiments continue until cumulative sum of differences remains within interval $\left(\ln B/d; \ln A/d\right)$. Hypothesis H is accepted once cumulative sum of differences is less than or equal to $\ln B/d$. Hypothesis H is rejected as soon as cumulative sum of differences is greater than or equal to $\ln A/d$. Sequential A/B testing implementations with preassigned values $p_1^0 = 0.1$, $p_2^0 = 0.12$, $\alpha = 0.05$, $\beta = 0.2$ and $p_1^0 = 0.1$, $p_2^0 = 0.13$, $\alpha = 0.05$, $\beta = 0.2$ are shown in Fig. 1, 2, respectively.

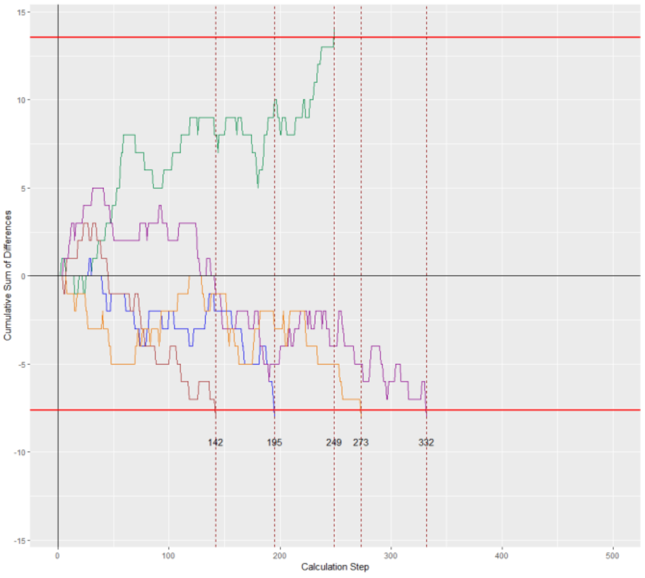

Fig. 1. Sequential A/B testing implementations I

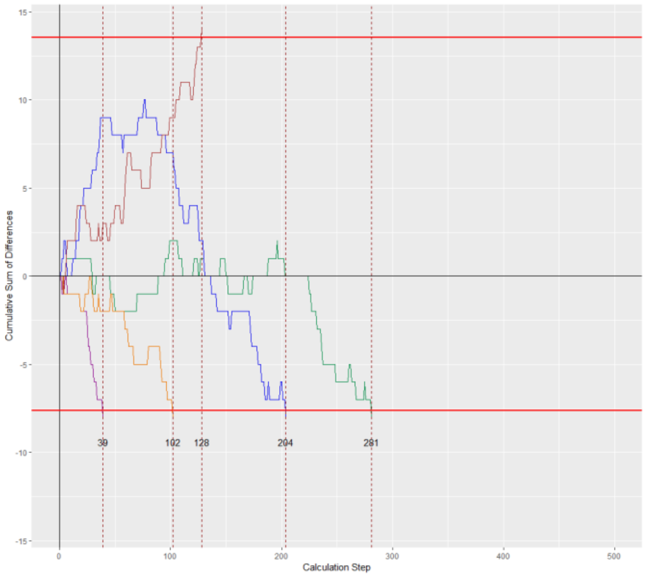

Fig. 2. Sequential A/B testing implementations II

B. Classical A/B Testing Implementation

Simple hypothesis $H: p_1 = p_2$ is formulated and two proportion Z-test is used for hypothesis testing.

Let $\hat{p}, \theta, \alpha, \beta$ be preassigned values. Confidence intervals for conversion rate difference $\hat{p}_1 - \hat{p}_2$ on each iteration are constructed. Hypothesis H is accepted if conversion rate difference isn't contained in the 95% confidence interval once the minimum number of unique visitors in one of the groups is equal to $n = 3532$. Hypothesis H is rejected if conversion rate difference is contained in the 95% confidence interval as soon as the minimum number of unique visitors in one of the groups is equal to $n = 3532$. Classical A/B testing implementations with preassigned values $\hat{p} = 0.1$, $\theta = 0.02$, $\alpha = 0.05$, $\beta = 0.2$, and $\alpha = 0.05$, $\beta = 0.2$, $\hat{p} = 0.1$, $\theta = 0.03$ are shown in Fig. 3, 4, respectively.

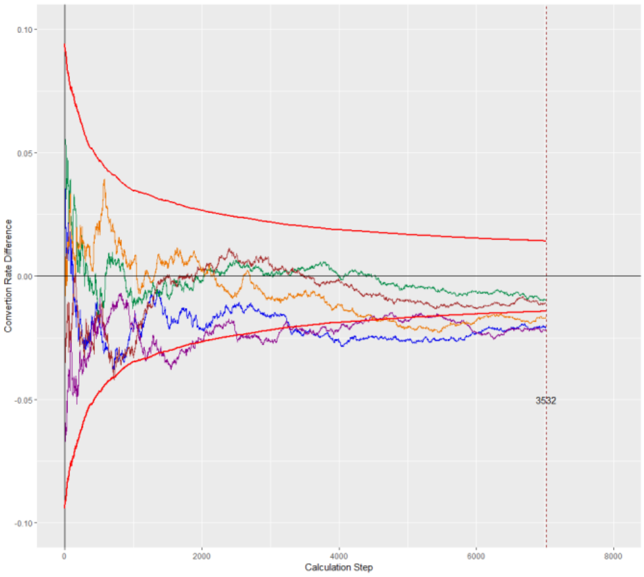

Fig. 3. Classical A/B testing implementations I

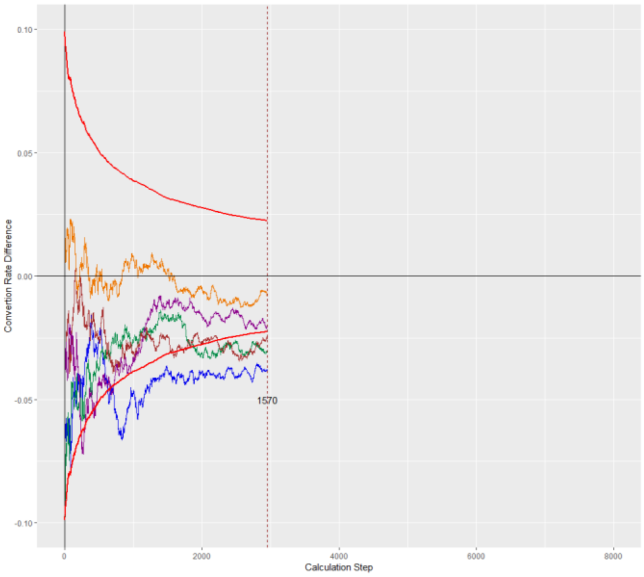

Fig. 4. Classical A/B testing implementations II

C. *Comparative Analysis*

For classical A/B testing with preassigned values p_1, p_2 and $\alpha = 0.05, \beta = 0.2$ the minimum number of unique visitors is calculated and presented in the third column of the Table I.

For sequential A/B testing with preassigned values p_1, p_2 and $\alpha = 0.05, \beta = 0.2$ mean number of visitors is calculated using 1000 implementations and presented in the fourth column of the Table I. This mean number is compared with the minimum number of unique visitors required for classical A/B testing. Results of comparison are presented in the fifth column of the Table I.

The advantages of sequential A/B testing are easy to see. As data collection can be terminated after fewer observations and decisions can be taken earlier, financial savings might be considerable. Sequential A/B testing allows to test different hypotheses for landing page optimization experiments and make changes based on actual data.

TABLE I. COMPARATIVE ANALYSIS

No.	***Preassigned values***	***Classical testing, No. of visitors***	***Sequential testing, mean No. of visitors***	***Improve-ment***
1	$p_1 = 0.1, p_2 = 0.12$	3532	337	90%
2	$p_1 = 0.1, p_2 = 0.13$	1570	175	88%
3	$p_1 = 0.1, p_2 = 0.14$	884	105	88%
4	$p_1 = 0.1, p_2 = 0.15$	566	70	87%
5	$p_1 = 0.1, p_2 = 0.16$	393	47	88%
6	$p_1 = 0.1, p_2 = 0.17$	289	36	87%
7	$p_1 = 0.1, p_2 = 0.18$	221	35	84%
8	$p_1 = 0.1, p_2 = 0.19$	175	31	82%
9	$p_1 = 0.1, p_2 = 0.2$	142	19	86%
10	$p_1 = 0.1, p_2 = 0.21$	117	18	84%
11	$p_1 = 0.1, p_2 = 0.22$	99	15	84%
12	$p_1 = 0.1, p_2 = 0.23$	84	14	83%
13	$p_1 = 0.1, p_2 = 0.24$	73	14	80%

We proposed a sequential A/B testing as the technique of decision making during A/B testing by sequentially collecting and processing the data. Sequential A/B testing procedure allows to perform landing page optimization experiments that reach conclusions 80% to 90% faster than classical A/B testing.

REFERENCES

[1] *Optipedia: The Optimization Glossary, Retrieved from http://www.optimizely.com/optimization-glossary/landing-page-optimization.html.*

[2] *R. Kohavi, R. M. Henne, and D. Sommerfield, Practical guide to controlled experiments on the web: listen to your customers not to the hippo, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 959–967, ACM, 2007.*

[3] *R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, Controlled experiments on the web: survey and practical guide, Data mining and knowledge discovery, vol. 18, no. 1, pp. 140–181, 2009.*

[4] *R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann, Online controlled experiments at large scale, Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1168–1176. ACM, 2013.*

[5] *R. Kohavi, A. Deng, R. Longbotham, and Y. Xu, Seven rules of thumb for web site experimenters, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014.*

[6] *A. Wald, Sequential tests of statistical hypotheses, Ann. Math. Statist., vol. 16, no. 2, pp. 117–186, 1945.*

[7] *L. Pekelis, D. Walsh, and R. Johari, The new stats engine, Technical report, Optimizely, 2015.*

[8] *R. Johari, P. Koomen, L. Pekelis, and D. Walsh, Peeking at a/b tests: Why it matters, and what to do about it, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1517–1525, 2017.*

[9] *R. Johari, L. Pekelis, and D. Walsh, Always valid inference: Bringing sequential analysis to A/B testing, arXiv preprint arXiv:1512.04922v3, Optimizely, 2019.*

[10] *E. Miller, Simple sequential a/b testing, Retrieved from http://www.evanmiller.org/sequential-ab-testing.html.*

[11] *M. Girshick, Contributions to the theory of sequential analysis. I, Ann. Math. Statist., vol. 17, no. 2, pp. 123–143, 1946.*