Paper 1070-2017

Cold-Start Solution to A/B Testing Using Adaptive Sample Size Modification

Bo Zhang, IBM; Liwei Wang, Pharmaceutical Product Development Inc.

ABSTRACT

A/B testing is a form of statistical hypothesis testing on two business options (A and B) to determine which is more effective in the modern Internet age. The challenge for startups or new product businesses leveraging A/B testing are two-fold: a small number of customers and poor understanding of their responses. This paper shows you how to use the IML and POWER procedures to deal with the reassessment of sample size for adaptive multiple business stage designs based on conditional power arguments, using the data observed at the previous business stage.

INTRODUCTION

In an A/B testing, an existing web page A (control) is modified to some degree to create a second version B (case) of the web page A. Later, half of the visitors are shown the control web page A and half are shown case web page B. The goal is to determine if version B performs better than version A for a given minimum relative change in conversion rate (effect size). In a traditional A/B testing design, a pre-determined number of visitors needs to be calculated based on the baseline conversion rate, the effect size, and the statistical significance. And then the A/B testing is stopped at the exact number of visitors pre-determined.

Such traditional A/B testing might not be suitable if we only have a very small number of visitors at the beginning of A/B testing and/or we have poor understanding of the conversion rates of our visitors. This motivates us to design an adaptive two-stage A/B testing. After a pre-determined number of visitors have been shown the two pages, the testing is paused, and the conversion rate is evaluated. By using the information at the end of the first stage of the A/B testing, the original assumption of the conversion rate can be re-evaluated in case it was too skeptical or too optimistic to the true conversion rate. The second stage of the A/B testing can be adjusted accordingly, while still controlling the Type I error rate.

TRADITIONAL A/B TESTING

A/B testing results in a dichotomous outcome. In other words, each visitor either responds or doesn't respond to the web page. For ease of exposition, throughout the paper, we will use the following notation:

- π_A: population response rate for web page A.
- π_B: population response rate for web page B.
- Δ: effect size, where $\Delta = \pi_B - \pi_A$.
- H_0: $\Delta = 0$: null hypothesis.
- H_1: $\Delta \neq 0$: alternative hypothesis.
- n_A: number of visitors assigned to web page A.
- n_B: number of visitors assigned to web page B.
- N: number of visitors assigned to web pages A and B, where $N = n_A + n_B$.
- α: the probability of a type I error, also called the level of significance.
- β: the probability of a type II error, where $1 - \beta$ is called power.

Here we combine the data into a summary test statistic:

$$T_n = \frac{p_B - p_A}{\sqrt{\bar{p}(1-\bar{p})(1/n_A + 1/n_B)}},$$

where

- $\bar{p} = \frac{X_A + X_B}{n_A + n_B}$.
- X_A: number of visitors responded to web page A, where $X_A \sim binomial\ (n_A, \pi_A)$.
- X_B: number of visitors responded to web page B, where $X_B \sim binomial\ (n_B, \pi_B)$.
- p_A: sample response rate for web page A, where $p_A = X_A/n_A$.
- p_B: sample response rate for web page B, where $p_B = X_B/n_B$.

With equal traffic allocation ($n_A = n_B = N/2$), the distribution of T_n under the alternative hypothesis H_1 follows a normal distribution with mean depending on N, π_A and Δ, with standard deviation depending on the parameters π_A and Δ.

Based on such distribution of T_n under the H_1, the sample size required for a two-side level- α test to attain at least $1 - \beta$ power to detect an increase of Δ, or greater, in the population response rate for web page B above the population response rate for web page A is

$$N = \frac{\left\{ Z_\beta \sqrt{\frac{\pi_A(1-\pi_A) + \pi_B(1-\pi_B)}{2(1-\bar{\pi})\bar{\pi}}} + Z_{\alpha/2} \right\}^2 4(1-\bar{\pi})\bar{\pi}}{\Delta^2},$$

where

- $\bar{\pi} = \frac{\pi_A + \pi_B}{2}$.
- $Z_{\alpha/2}$ denotes the $(1 - \alpha/2)$-th quantile of a standard normal distribution.
- Z_β denotes the $(1 - \beta)$-th quantile of a standard normal distribution.

Considering an A/B testing example where $\pi_A = 0.03, \Delta = 0.006, \alpha = 0.10$ and $\beta = 0.15$, we can use PROC POWER of SAS® for the sample size calculation as shown below:

```
proc power;
    twosamplefreq TEST=pchq
    proportiondiff = 0.006
    refproportion = 0.03
    npergroup = .
    power = 0.85
    alpha = 0.10;
    title "Sample Size Calculation for Traditional A/B Testing";
run;
```

The output is produced in Figure 1. As you can see, in order to achieve 85% power to detect at least 0.6% improvement in response rate, the A/B testing needs 12744 visitors per variation of the web page. This traditional A/B testing design has limitations. There is uncertainty as to what Δ it is appropriate to use. For example, in the above test, there will be debate on why 0.006 is being used. Hence, there is a need to re-calculate the sample size by leveraging the results obtained in the first stage of the test.

Sample Size Calculation for Traditional A/B Testing

The POWER Procedure
Pearson Chi-square Test for Two Proportions

Fixed Scenario Elements	
Distribution	Asymptotic normal
Method	Normal approximation
Alpha	0.1
Reference (Group 1) Proportion	0.03
Proportion Difference	0.006
Nominal Power	0.85
Number of Sides	2
Null Proportion Difference	0

Computed N Per Group	
Actual Power	N Per Group
0.850	12744

Figure 1. Sample Size Calculation for Traditional A/B Testing

ADAPTIVE TWO-STAGE A/B TESTING

The challenge for startups or new product businesses leveraging A/B testing are two-fold: a small number of customers and poor understanding of their responses. The adaptive two-stage A/B testing starts with a "small but able to handle" pre-determined number of visitors. Additional visitors need to be involved to the A/B testing only if you obtain the equivocal results from the first stage.

The single metric to evaluate to make decisions on whether to increase the sample size is conditional power, denoted by CP.

If the conditional power is too low and less than a pre-computed lower limit CP_{lower}, then the testing will be considered as ineffective and will be stopped. If conditional power is too high and larger than the originally planned power $1-\beta$, then the results are favorable and the A/B testing can be continued to the originally planned number of visitors. There is no need to increase the sample size. If the conditional power is neither too low nor too high, the results are considered as equivocal and the sample size needs to be increased to the boost the power to $1-\beta$. To summarize, the adaptive two-stage A/B testing partition the CP into three regions:

- Unfavorable: $CP < CP_{lower}$. The testing will be considered as ineffective and will be stopped.
- Equivocal: $CP_{lower} \leq CP < 1-\beta$. The results are considered as equivocal and the sample size needs to be increased to the boost the power to $1-\beta$.
- Favorable: $CP \geq 1-\beta$. The results are favorable and the A/B testing can be continued to the

originally planned number of visitors.

The process of adaptive two-stage A/B testing can be described in Figure 2.

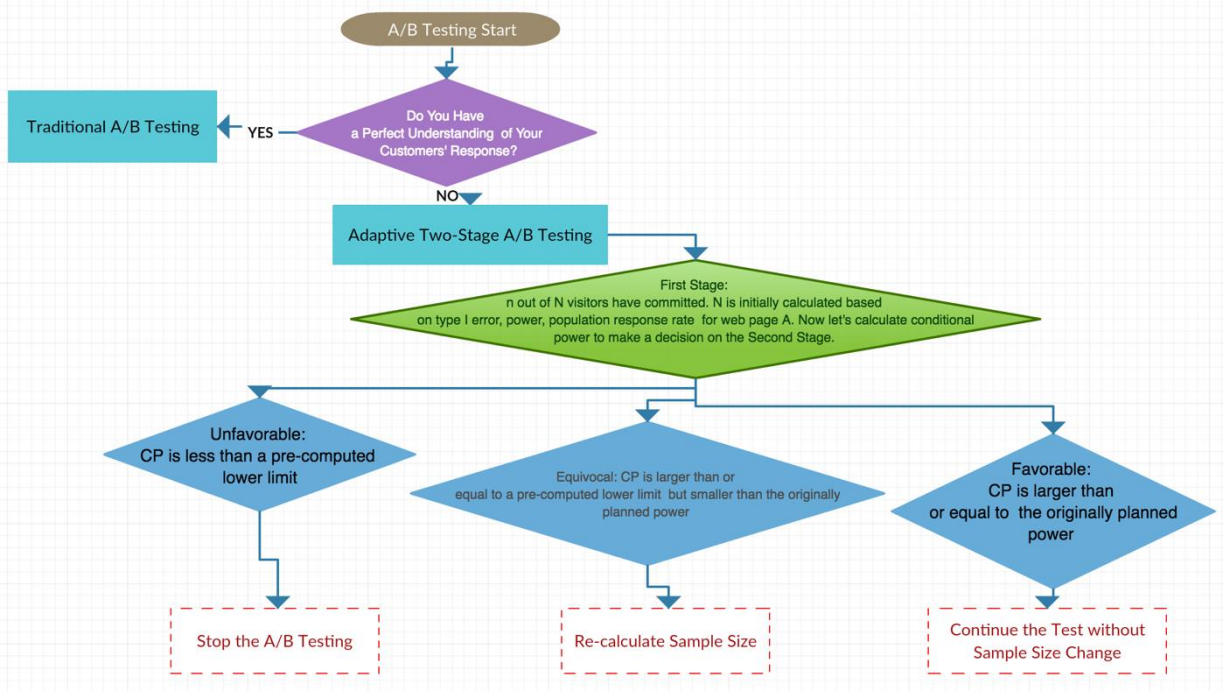

Figure 2. Adaptive Two-stage A/B Testing Flowchart

Before we jump into the details of each step, let's first introduce the following additional notations:

- n_{A1}: number of visitors assigned to web page A for the first stage.
- n_{B1}: number of visitors assigned to web page B for the first stage.
- n: number of visitors assigned to web pages A and B for the first stage, where $n = n_{A1} + n_{B1}$
- T_n: the observed test statistics at the end of first stage when n visitors have been involved.

LOWER LIMIT OF CONDITIONAL POWER

The pre-determined lower limit of conditional power should not be too small to result in an inflated type I error. Mehta an Pocock (2011) showed that the minimum CP_{lower} accepted depends on n/N, the significance level α, and the originally planned power $1-\beta$. Here, we set $\alpha = 0.10, \beta = 0.15, \text{and } n/N = 0.5$. Based on the Lemma 1 in Mehta an Pocock (2011), we can use PROC IML of SAS for the gridding association between critical boundary and conditional power as shown below:

```
proc iml;
    cp=t((10:90)/100);
    alpha=0.1;
    beta=0.15;
    power=1-beta;
    zalpha=quantile("Normal",1-alpha/2,0,1);
    zbeta=quantile("Normal",power,0,1);
    n=12744;
    Np=12744*2;
    t=n/Np;
    zt=(sqrt(t))*(quantile("Normal",cp,0,1)*sqrt(1-t)+zalpha);
    n2_0=n+(n/zt##2)#((zalpha*sqrt(Np)-zt#sqrt(n))/sqrt(Np-n)+zbeta)##2;
    n2=n2_0;
```

```
    n2=ceil(n2_0);
    n2[loc(n2_0<np)]=np;
    n2star=n2-n;
    b=(sqrt(n2star/(np-n))#(zalpha#sqrt(np)-
    zt#sqrt(n))+zt#sqrt(n))/sqrt(n2);
    create out var {cp zt n2 b zalpha power};
    append;
    close out;
quit;
```

Then based on the equation (12) in Mehta an Pocock (2011), we can use PROC SQL and PROC GPLOT of SAS for the determination of the minimum CP_{lower} as shown below:

```
proc sql noprint;
    select zalpha into: zalpha from out
    where zalpha >.
    ;
    select power into: power from out
    where power >.
    ;
    select min(cp) into: CPlower from out
    where b<=&zalpha
    ;
quit;

ods graphics on;
symbol1 interpol=join value=dot;
proc gplot data=out;
    plot b*cp/vref=&zalpha href=&CPlower &power;
    title "b versus CP";
    title2 "Promissing Zone Requires CP Such That b<=z_alpha";
    title3 "CPlower = %cmpres(&CPlower), Target Power = %cmpres(&power)";
run;
```

The calculated lower limit of conditional power is 0.31. And the output plot is shown in Figure 3.

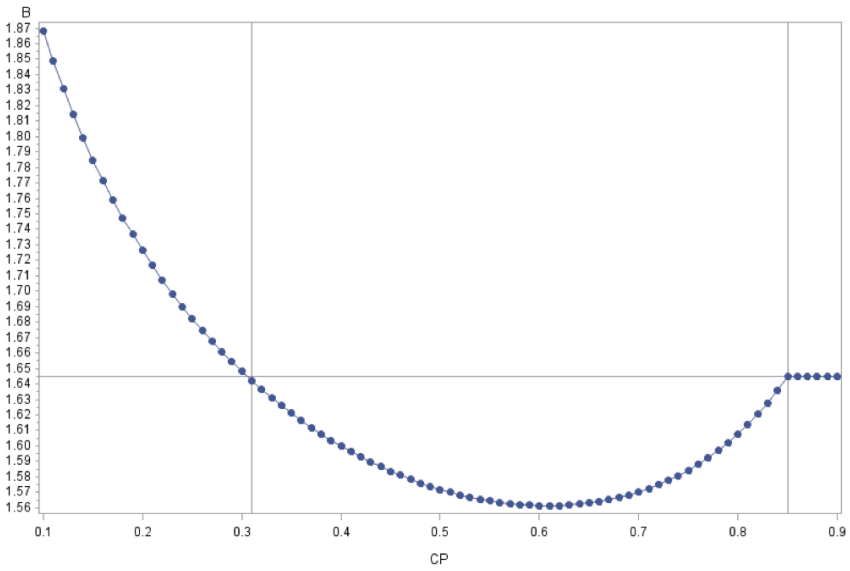

Figure 3. Critical Boundary and Conditional Power

CONDITIONAL POWER

In Lan and Wittes (1988), assuming the second stage data will be similar to what has been observed in the first stage, the conditional power given the n samples in test is computed using the following formula:

$$CP = \Phi\left(\frac{\frac{T_n}{\sqrt{n/N}} - Z_{\frac{\alpha}{2}}}{\sqrt{1-\frac{n}{N}}}\right).$$

Here we set $\alpha = 0.10, \beta = 0.85, \mathrm{n/N} = 0.5$, $X_{A1} = 191$, $X_{B1} = 216$, $n_{A1} - X_{A1} = 6181, and\ \ n - X_{B1} = 6156$. In other words, among all the visitors assigned to web page A in the first stage, 191 responded and 6181 didn't response. Among all the visitors assigned to web page B in the first stage, 216 responded and 6156 didn't response. Then we use PROC IML of SAS for the conditional power calculation as shown below:

```
data test;
   input group $ response count;
   cards;
   A  1 191
   A  0 6181
   B  1 216
   B  0 6156
   ;
run;

ods output ChiSq=ChiSq;
proc freq data=test;
   tables group*response/chisq;
   weight count;
   title "Observed Test Statistics at End of First Stage";
run;

proc sql noprint;
   select sqrt(Value) into :z
   from ChiSq
   where Statistic="Chi-Square"
   ;
   select sum(count) into :n
   from test
   ;
quit;

proc iml;
   alpha=0.1;
   beta=0.85;
   zalpha=quantile("Normal",1-alpha/2,0,1);
   zbeta=quantile("Normal",beta,0,1);
   n=&n;z=&z;
   Np=12744*2;
   t=n/Np;
   title "Conditional Power Based on Observed Test Statistics";
   *calculate conditional power;
   zz=(&z/sqrt(t)-zalpha)/sqrt(1-t);
   cp=round(cdf("Normal",zz,0,1),.001);
   print cp;
```

```
    *recalculate sample size;
    NZn=n+(n/(z**2))*(((zalpha*sqrt(Np)-z*sqrt(N))/sqrt(Np-n)+zbeta)**2);
    *Round up re-calculated sample size and make it an even number;
    N2=ceil(NZn/2)*2;
    create ReCal var {alpha beta zalpha zbeta z cp n2};
    append;
    close ReCal;
quit;
```

The outputs are shown in Figure 4, Figure 5, and Figure 6.

Observed Test Statistics at End of First Stage

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of group by response		
	response		
group	0	1	Total
A	6181 48.50 97.00 50.10	191 1.50 3.00 46.93	6372 50.00
B	6156 48.31 96.61 49.90	216 1.69 3.39 53.07	6372 50.00
Total	12337 96.81	407 3.19	12744 100.00

Figure 4. Frequency Table

Statistics for Table of group by response

Statistic	DF	Value	Prob
Chi-Square	1	1.5863	0.2079
Likelihood Ratio Chi-Square	1	1.5873	0.2077
Continuity Adj. Chi-Square	1	1.4619	0.2266
Mantel-Haenszel Chi-Square	1	1.5862	0.2079
Phi Coefficient		0.0112	
Contingency Coefficient		0.0112	
Cramer's V		0.0112	

Figure 5. Test Statistics

Conditional Power Based on Observed Test Statistics

cp
0.576

Figure 6. Calculated Conditional Power

SAMPLE SIZE RE-CALCULATION

As we can see from the above step that the calculated conditional power (0.576) is between lower limit (0.31) and the originally planned power $1-\beta$ (0.85), the adjustment of the sample size is then needed. Using equations in Mehta an Pocock (2011), the re-calculated sample size N_{new} will be

$$N_{new} = n + \left(\frac{n}{T_n^2}\right)\left(\frac{Z_{\frac{\alpha}{2}}\sqrt{N} - T_n\sqrt{n}}{\sqrt{N}-n} + Z_\beta\right)^2.$$

The sample size re-calculation is already implemented when we calculate the conditional power using PROC IML of SAS from the above step. Here we just put all the key metrics together using:

```
data decision;
    set ReCal;
    length decision $100.;
    if cp<&CPlower then decision="Stop A/B Testing";
    else if &CPlower<=cp<&beta
    then decision="Increase Total Sample Size to"||n2;
    else decision="Continue A/B Testing to Originally Planned Sample Size";
    label alpha="Significance Level"
          beta="Planned Power"
          z="Test Statistics at End of First Stage"
          cp="Conditional Power"
          n2="Re-calculated Sample Size"
          decision="Decision"
    ;
 run;

proc print data=decision label noobs;
    var alpha beta z cp decision;
    title "Decision Made at the End of First Stage";
run;
```

The output is shown in Figure 7.

Decision Made at the End of First Stage

Significance Level	Planned Power	Test Statistics at End of First Stage	Conditional Power	Decision
0.1	0.85	1.25948	0.576	Increase Total Sample Size to 48280

Figure 7. Final Decision

CONCLUSION

Traditional A/B testing completes with a hypothesis, a control web page and a variation web page, a calculated sample size, and a statistically calculated result. The adaptive two stage A/B testing discussed in this paper offers additional benefits to the traditional A/B testing: to handle cold-start issue and to best facilitate business decision making. Based on the calculated conditional power and lower limit, we can decide if there is a need to increase the sample size, to stop the A/B testing or to continue.

REFERENCES

Lan, K. G., & Wittes, J. (1988). The B-value: a tool for monitoring data. *Biometrics*, 579-585.

Mehta, C. R., & Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in medicine*, *30*(28), 3267-3284.

Chung, St. Clare Groulx, Adrienne Moon, Kyung-hee (Kelly). (2007). Using SAS to Determine Sample Sizes for Traditional Two-Stage and Adaptive Two-Stage Phase II Cancer Clinical Trial Designs. SAS Global Forum Proceedings.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bo Zhang
IBM
Email: bozhang@us.ibm.com