Portfolio: Web Analytics Case Study

Mitchell D. Shuey

January 17, 2019

# About the Data:

Several variations of Montana State University’s library webpage were made in the pursuit of A/B Testing their front page. Specifically, the altered piece of the website was the headline linking to user services the library provides. The original, “Interact”, had a noticably small click rate. Alternatives were considered:

“Connect”, “Learn”, “Help”, and “Services”.

Google Analytics along with a click-tracker called CrazyEgg were used to generate the data. Users, when visiting the site, were randomly sent to one of the five variations. Our goal is to determine which led to more clicks (the “conversion rate” or click-through rate) overall. If led to a purchase, we could also calculate average revenue-per-click.

The first dataset, titled Lib\_msu, shows all the pages of the whole website that were clicked on between May 29th and June 18th, 2013 (the length of the trial).

The next five, ctrl and v1-v4, are the home pages’ click counts over those three weeks for each variation. The code importing these files was hidden because of the room it takes up.

## Warning: package 'tidyverse' was built under R version 3.5.2

## -- Attaching packages ------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## Warning: package 'ggplot2' was built under R version 3.5.1

## Warning: package 'tidyr' was built under R version 3.5.1

## Warning: package 'dplyr' was built under R version 3.5.1

## -- Conflicts ---------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

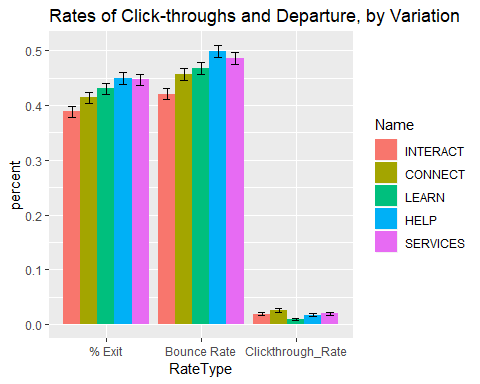
Now we’ll begin taking apart the data. The piece of interest is the clicks on each subtitle, compared to the total overall from the Lib\_msu file.

clicks = rbind(ctrl[10,], v1[7,], v2[11,], v3[8,], v4[8,])  
index = Lib\_msu[3:7,]  
#Re-ordering  
index=index[c(4,5,2,3,1),]  
clicks=cbind(clicks[,-6], index)  
#Calculating click rate  
clicks$Clickthrough\_Rate=clicks$`No..clicks`/clicks$Pageviews  
clicks$ctr\_share=clicks$No..clicks/sum(clicks$No..clicks)  
clicks

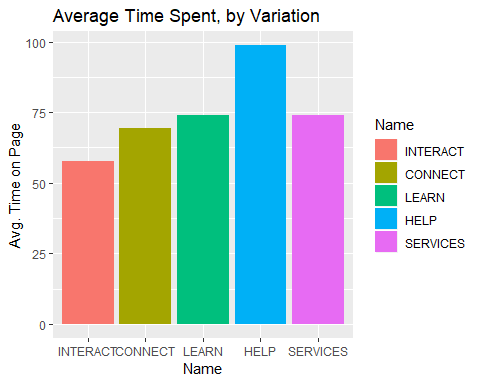
## Element.ID Tag.name Name No..clicks Visible. Page Pageviews  
## 10 87 a INTERACT 42 true / 2206  
## 7 92 a CONNECT 53 true /index2.php 2099  
## 11 87 a LEARN 21 true /index3.php 2213  
## 8 92 a HELP 38 true /index4.php 2210  
## 81 87 a SERVICES 45 true /index5.php 2284  
## Unique Pageviews Avg. Time on Page Entrances Bounce Rate % Exit  
## 10 1750 57.83543 1694 0.4208973 0.3884859  
## 7 1618 69.40000 1583 0.4560960 0.4140067  
## 11 1727 74.00000 1663 0.4678292 0.4306371  
## 8 1701 98.77056 1657 0.4990947 0.4497738  
## 81 1806 73.93903 1759 0.4860716 0.4470228  
## Page Value Clickthrough\_Rate ctr\_share  
## 10 0 0.019038985 0.2110553  
## 7 0 0.025250119 0.2663317  
## 11 0 0.009489381 0.1055276  
## 8 0 0.017194570 0.1909548  
## 81 0 0.019702277 0.2261307

We see that with a rate of 2.5%, “Connect” has the highest click-through rate, but we need to determine the significance of this with a test. Visuals including error bars will also help judge the differences of results.

library(ggplot2)  
plotset=gather(clicks, RateType, percent, 11, 12, 14)  
#Calculating standard deviations for each rate versus the variation's , bernuolli  
plotset$sd=sqrt(plotset$percent\*(1-plotset$percent)/plotset$Pageviews)  
#plotting  
rates=ggplot(plotset, aes(RateType, percent, fill=Name))  
bars=rates+geom\_col(position="dodge")  
 bars+ geom\_errorbar(aes(ymin=percent-sd, ymax=percent+sd), width=.4, position=position\_dodge(.9))+ggtitle("Rates of Click-throughs and Departure, by Variation")



ggplot(plotset, aes(Name, `Avg. Time on Page`, fill=Name))+  
 geom\_col(position="dodge")+ggtitle("Average Time Spent, by Variation")



share=ggplot(clicks, aes(No..clicks, fill=Name))

Exit Rate- Leaving the site entirely after reaching one of the home page variations. May have visited other pages prior to ending at this page. Bounce Rate- Leaving the site after reaching one of the home page variations. Entering and leaving, nothing more.

While it’s hard to interperet whether a higher or lower exit rate means users found what they were looking for, we know that a lower bounce rate is better because if the users found what they were looking for

# Hypothesis testing

We initially assume the null: All treatments (versions) have an equal conversion rate The alternative: they are unequal in some degree.

We can then dive into comparing all five click-through rates against each other, with the p-values in the matrix representing the chance that the difference between, say, the control (1) and “Help” (4) was due to the randomness of the data:

prop.test(clicks$No..clicks, clicks$Pageviews)

##   
## 5-sample test for equality of proportions without continuity  
## correction  
##   
## data: clicks$No..clicks out of clicks$Pageviews  
## X-squared = 15.836, df = 4, p-value = 0.003248  
## alternative hypothesis: two.sided  
## sample estimates:  
## prop 1 prop 2 prop 3 prop 4 prop 5   
## 0.019038985 0.025250119 0.009489381 0.017194570 0.019702277

#The p-value is significant enough to plan further testing  
  
pairwise.prop.test(clicks$No..clicks, clicks$Pageviews, p.adjust.method = "bonferroni")

##   
## Pairwise comparisons using Pairwise comparison of proportions   
##   
## data: clicks$No..clicks out of clicks$Pageviews   
##   
## 1 2 3 4   
## 2 1.0000 - - -   
## 3 0.1075 0.0011 - -   
## 4 1.0000 0.8322 0.3553 -   
## 5 1.0000 1.0000 0.0646 1.0000  
##   
## P value adjustment method: bonferroni

# Conclusion

According to our data, despite the Click-through rates being different, using the standard 5% p-value cutoff, few have a significant difference from other variations 1-on-1. The results of the original research state that “Learn” has the worst click-through performance. I conclude no significant difference was found within the data here; however this does not mean we are unable to make a recommendation.

Looking at the average time chart, we can see that the control, “Interact”, has the shortest average time on page. In a library context, where finding what you want on the page quickly is important, this could mean that the experiment itself led students to take more time on the home page to make sure they were going to select what they needed. They might be used to looking for “Interact”, versus interpereting the new variations. This is an important effect to consider when changing sections of a website.

Given the chance to repeat the test, less variations, or limiting testing to one alternative at a time would be a good idea. It’s worth noting that comparing all of these proportions to each other at once calls for adjustments to the p-value. In this case, using the “bonferroni” method, all p-values were multiplied by 15.

It’s worth noting that on the MSU library’s current page, the choice was made to use “Services”. The original paper made different choices in calculating click rates and decided to forego testing for significance; I have emailed the author and asked about the calculation methodology. The original paper also conducted user surveys, which offer far more contextual insight into a user’s behavior and final click decisions. A good insight campaign uses all available sources of consumer experience information.

Works Cited: Young SWH (2014) A/B Testing Web Analytics Data [dataset]. Montana State University ScholarWorks. <http://doi.org/10.15788/m2rp42>