University of Virginia

Final Project on A/B Testing

Maria Aguilera, Kelsey Moon, and Michelle Joung

STAT 4220: Applied Analytics for Business

Professor Ferrara

April 22, 2019

**Table of Contents**

***Background of Statistical Analysis.……………………..…………………………..……...…....3***

***Business Questions and Objectives***

***Business Understanding…………………..…………………………………….…..…....5***

***Data Background***

***Data Understanding....………………………..…………………………………….....….6***

***Data Preparation………………………………………………………………………….6***

***Data Visualizations & EDA..……………………………………………………….....….7***

***Results & Discussions***

***Assumptions...……………………………………..…………………………………..….8***

***Evaluation...……………………………………..……………………………..………....8***

***Recommendations...……………………………………..………………...……….……..9***

***Works Cited.…………………………………….……………...…..……………………...…….11***

***Appendix....…………...……………………………………………..…………………..……….12***

## 

***Background of Statistical Analysis***

The method that we will be highlighting in this report is A/B testing. One can think of this method as a way of primarily testing out a new idea, generating a hypothesis in order to carry out an experiment, utilizing statistics to validate or fail to validate said hypothesis, updating to the more successful idea, and repeating the cycle if necessary. The variety of tests that can be performed with A/B testing are endless and some examples include testing different brands of clothing, coffee or even chairs, but it is mainly associated with comparing two versions of a web page or application. For example, a question an entrepreneur may have is whether or not changing the logo of her company on her website to color instead of black and white would improve conversion rates. A conversion rate is defined as the proportion of consumers who perform an action beneficial to the company divided by the number of consumers who visited the website overall (Piccinini, 2018). An action can be the time a consumer spends on a website, the engagement of a consumer such as clicking a ‘like’ button, or tracking the drop off rate which is seeing if a visitor continues to the next page (Piccinini, 2018). The action is usually denoted as 1 if it is followed through and 0 otherwise. Now, the control and test in the experiment must be defined. In our example, the control would be keeping the old page which is the monochrome one and the test would be the new page which is in color.

After deciding on the purpose for employing the method, and defining the test and control variables, a company or entrepreneur should select the duration of the A/B test. One can compare the conversion rates between a control and test set weekly, biweekly, monthly or even annually but it is important to note that the longer timeframes will ultimately lead to slower results. A power analysis can be run to determine the length of an experiment, define the power (1 - Beta) which is the probability of correctly rejecting the null hypothesis, determine the significant level (alpha), and effect size which is the difference between the control and experimental group population’s averaged means divided by the assumed common popular standard deviation (Piccinini, 2018). Higher power, lower significance level, and smaller effect size will often result in more data points. It is crucial that the control and tests are run simultaneously as well as during the same timeframe. For example, a control conversion rate in March could be 0.4 compared to August which could be 0.9. When running the test set, if it is in March it should be contrasted to the control set in March since comparing it to the higher control August rate would be an inaccurate comparison; seasonal variation must be factored. To further demonstrate the method, a hypothesis in our case could be “If the original monochrome logo on the company website is changed to color, it will lead to higher conversion rates.” The null hypothesis would be that there is no significant difference in conversion rates between the control and test whereas the alternative would say there is. The primary objective of the test in this scenario is to analyze whether the updated version of the website led to increases in conversion rates. If the entrepreneur does not have data collected for the test set which is more often the instance, statistical software such as R can be used to help determine the sample size needed depending on the percentage increase desired such as 5 or 10 percent. Generally, the lower the increase in rate, the more data points will be needed (Piccinini, 2018). This concludes the necessary components needed before analyzing the results.

The entrepreneur can now run tests such as logistic regression for a binary categorical independent variable, a t-test or linear regression for a continuous dependent variables depending on the size, and create visualizations such as line graphs in order to compare the conversion rates between the two groups. Once the results are summarized, the p-value, estimate, and standard error can be used to either reject or fail to reject the null hypothesis. If the null is rejected, the entrepreneur will know that the change is statistically significant and should be implemented right away. Furthermore, follow-up experiments can be conducted in order to improve conversion rates even more. Examples could include changing the background picture of the webpage, the font of the title, or different words and how the amount of load time could be affected by it. However, it should be stressed that the changes should be small and each experiment should be conducted separately in order to accurately address which specific change accounted for fluctuating conversion rates. Changing multiple variables could lead to a side effect which is an unintended consequence of a change such as the size of the header and what a visitor first sees without scrolling down the page which could lead to them leaving or staying. When more than one independent variable is tested though, a multivariate test is possible but it is far more complex and not recommended. Confounding variables are also a significant part of A/B testing and this is ‘an element of the environment that could affect your ability to find out the truth of the experiment’ and can be either internal or external (Piccinini, 2018). Internal accounts for things such as the number or frequency of words in a title rather than the words themselves whereas external factors could include the demographic of the people who visited or if a popular movie related to the website was just released which could account for an influx of visitors. Confounding variables are one of the disadvantages of A/B Testing, this as well as possible season variation issues, the implementation of the method taking a long time, and possible false positive results may misconstrue the conclusions and lead to inaccurate results.

An assumption of A/B Testing that must be talked about is the difference between within and between groups. Within groups consist of participants who are able to see both the control and test experiment and you can analyze which one people find more attractive. This group will lead to a higher power since they have the ability to contrast the differences side by side. The between group consists of two groups where one sees the control and another sees just the test. This is ideal if visitors frequent the site and when creating this group, the assumption is that there should be no qualitative differences between the two participant groups and they should come from the same random group of all possible participants (Piccinini, 2018). Additionally, we have stressed A/B testing which is comparing two a control and a test condition since this is the method we are particularly interested in, but there are two other types. First, there is A/A testing which is when there are two groups of control conditions. For example, group 1 can see the control and participate in the experiment and then group 2 will do the same. Another test is A/B/N testing which is comparing a control condition to any number of different test conditions. This was previously mentioned when an entrepreneur wants to test multiple independent variables, but the statistics are far more complicated and the amount of data points needed would be multiplied immensely in order to be certain that there was an effect.

A/B testing is a continuous cycle of updating variants of a design in order to improve the desires of entrepreneurs or companies. When all the necessary steps are taken and analyzed correctly, depending on the specific type of test, it can optimize devices or websites and ultimately lead to an increase in profits. The advantages as well as the application are endless whether its increasing the rate of successful bookings on travel websites or increasing check out orders for a clothing website. The technique is further analyzed in this paper through a real-world example of an E-commerce company.

**Possible Business Question(s) or Objective(s)**

***Business Understanding***

Our primary business objective is to examine our data through an A/B Testing method to contrast two versions of a webpage, evaluate the performance of each, and decide which would benefit the company. The data we are examining contains information about an E-commerce company website and includes conversion rates of an old page compared to an updated one (Zhang, 2018). Since companies do not want to publicize their marketing strategies, the data is quite obscure. We are not given the actual webpage, what type of product the company is promoting, or any website features so our business objectives are limited. If we were given more detailed information, possible questions that may be asked are whether specific changes in the old page such as the color, wording of titles or locations of tabs within the website allow for higher conversion rates, or whether the rates remain unchanged.

A major benefit of the A/B testing method is that it may be utilized to help E-commerce companies determine what changes will increase the user conversion rate. This in turn could lead to more purchases of products which would increase the company’s revenue. The results of the analysis may also be used to determine whether or not the company needs to invest on an updated website design which could be a potential cost. Another benefit is that A/B testing can be used on a variety of different independent variables so the test can be repeated to further observe what increases conversion rates. Within our data, the A/B testing method helps to identify which specific alteration of the website can maximize the conversion rate thus enhancing the quality of the website as well as the company’s profit.

The biggest data quality issue we face is the fact that we do not have sufficient descriptions of the actual collection of the data. In addition, this lack of information leads us to make educated guesses at best which can potentially lead to making inaccurate conclusions. Apart from the given five variables, of which three give the exactly same information, we are also unclear as to how they were measured. Additionally, there is no information about what constitutes as a conversion, so we are assuming it was defined as a consumer clicking on an associated subpage and or link. There could have been a multitude of different ways of measuring this depending on the company’s goal. For example, it could have been completing purchases or even clicking a ‘like’ button so having limited information is quite dangerous. We are also unaware of the layout of the webpage and it is difficult to imagine one since the factors that contribute to the design of the webpage are unlimited. Different colors border lines of the images, layout or font and other factors could have all contributed to the conversion rate. Moreover, there is no information on the demographic of the website visitors, their reasoning behind coming to the website and if there was a purpose to it. In addition, there could have been repeat visitors which is a factor to consider since frequent visitors often show momentum behavior. Momentum behavior is characterized by the routine patterns visitors take when visiting a website. The patterns are based on a known design, therefore when a new website design is implemented, the returning visitor may become less frequent since they prefer and are used to the older/first found web page; thereafter impacting conversion rates (Shukairy, 2017).

Furthermore, possible technical errors on the website during the data collection period may have altered the conversion rate of the new and old pages, essentially impacting our conclusions on the data after analyzation. A specific example of a technical error is the Flicker effect, which is common and describes when an original page is briefly displayed before the alternative appears during an A/B test (Brebion, 2018). In our scenario, the original page would be described as the old page and the alternative as the newer page with modifications. Due to these issues, we may face a lot of difficulties in interpreting the results of our analysis as well as further research since the problems have not been noted. In addition, having the data already analyzed on kaggle instead of us running the experiment and collecting data was also another potential issue since we were not able to control the sample size or the type of variables. If we were in charge of collecting our data, we could have had more specific variables such as what word or picture a consumer clicked on or what product they purchased and the total amount of it. On the bright side, we found that our data did not contain any missing values within our variables/data.

**Data Background**

***Data Understanding***

The data analyzed was collected from Kaggle, and is a data frame named *“A/B Testing”* published by contributor Luyuan Zhang. Within the dataframe, 294,478 observations were collected and there were no missing variables. The data had five different variables which are noted as user id, time stamp of the website visit, group, landing page and converted page. The user id is specific to each consumer so there were a couple repeats throughout. The timestamp variable allows for the disclosure of the time of collection for this data frame, which is from January 2nd, 2017 to January 24th, 2017. The variable group identified whether each observation was a control group or a treatment group. The variable landing page is subdivided into old page and new page and ideally old would correlate with control and new with treatment. The variable converted page has 2 possible values of 0 and 1. The value 1 indicates that the website page results in a conversion while the value 0 indicates that the website page fails to result in a conversion. As seen by figure 1 and figure 5, there are 17,489 old web pages that resulted in a conversion, 127,785 old pages that did not result in a conversion, 17,264 new web pages that resulted in a conversion and 129,741 new web pages that did not result in a conversion. Based on just this rudimentary analysis, the rates between each group are eerily similar.

***Data Preparation***

There weren’t any missing data values or outliers in the original data. However, we identified multiple data values that did not align correctly to the “control” and “treatment” group. For instance, when the new page was tested, it was recorded as a control group when it should have been labeled as the treatment group. A total of 3,893 observations were not aligned, and thus removed from the cleaned data set. We also sorted each user id by the time frame and only subsetted the first recorded instance of each user to remove any potential repetition. In order to see how the conversion rates varied by day during the collection period, we formatted the data into year, month and day. Then, we subsetted just the day which ranged from 2 to 24 of the month of January since this would let us analyze the data more in depth.

There are many limitations in our data such as the lack of information about what exactly constitutes as a conversion. The time period of data is also limited since it only shows the conversion rates for January. It might not be entirely statistically accurate to generalize the overall website’s conversion rates based on one month’s data due to seasonal variations. However, as the data collection period is more than 3 weeks, and both the test and control set are collected during the same time, we are assuming that the company was particularly interested in this month.

***Data Visualizations & Exploratory Data Analysis***

To start off our exploratory data analysis, we created a table of means as well as a table of standard deviation between the groups. For the new and old page, the means were 0.1188 and 0.1203863 and the standard deviations were 0.3236 and 0.3254, respectively. Based on even these preliminary results, there is almost no difference between the two groups. There were no outliers or missing values in our data set so we were able to find the distinct conversion rates right away. Referring to table 1, you can see that three of our variables are binary, user id is irrelevant for analysis, since it denotes identification, and the timeframe is limited so most of our EDA was merely comparing the conversion rates to each variable.

The overall conversion rate which divides the converted rows by the total dataframe is 0.119 and the bar chart in Figure 4 shows the difference within each group. As you can see, there is a minimal almost nonexistent difference in the rate between the old and new page which is difficult to see even within the bar chart. Due to this, we thought that creating conversion and non-conversion rate tables would be more beneficial so we could quantitatively see the difference between the new and old page. After subsetting the four combinations between the converted and landing page columns, Figure 1 was created. Looking at the conversion rate column of Figure 1, the difference between the old and new page is merely 0.16%. The 12.04% rate for the old page is known as the conversion rate control, and the 11.88% for the new page is referred to as the conversion rate treatment value.

The data was collected from January 2nd, 2017 to January 24th, 2017. To visualize if there was a difference between the conversion rates of the old and new page each day, Figure 3 was created. This visualization is a line graph with the color red showing the old page and blue showing the new one. Both lines are extremely sporadic and alternate having the higher conversion rate, but it is important to note that this is due to the y-axis scale showing a range of just 1%. Had the line been stretched out horizontally, there would be almost no difference between the two groups and the figure also shows that there are no distinct, constant patterns with the conversion rate over the three week period.

Figure 3 might have been aesthetically pleasing, but Figure 5 with the comparison of the two bar charts is able to give more of a quantitative visualization. When looking at the conversion rates for both the old and new page it is relatively low compared to those who did not convert. The two bar charts are almost identical in length showing that the new page did not seem to improve the conversion rates.

No cross validation was needed for A/B testing since our dataset is not focused on machine learning so there was no need to find an optimal value in this case.

**Results & Discussions**

Due to our dataset lacking pertinent information, we have decided to base this section off of a made-up scenario. We will assume that the E-commerce website was testing two versions of an advertisement and the conversion rate measured depicted whether a consumer clicked on the associated link or not. The old advertisement is the control whereas the remodeled one is the treatment and our null hypothesis is that there is no difference in conversion rates between the two adverts.

***Assumptions***

Assumptions that need to be satisfied in order for the testing method to be implemented were about the difference between the Between Groups, and the necessary assumptions for the statistical test. For our case, since our dependent variable, conversion rate, is a categorical one, we decided to use a chi-square test to see if there was a difference between the old and new page. Firstly, an assumption of the chi-square test is that it must be applied on data sets that are large enough in terms of sample size. Secondly, another assumption of the chi-square test is that it cannot be used on correlated data which means that the observations need to be independent (Deshpande, 2019). For the between group assumptions, if there were two distinct groups, one who saw the test and the other that saw the control, they should be coming from the same random group and there should be no qualitative differences between the groups. Since we did not collect the data we are unable to confirm that this assumption is met, but since the users all have different ids, were extracted based purely on the time, and our sample size is large, we are assuming that it was based off the dates chosen and nothing else. For the chi-square assumptions, our data set holds over 30,000 observations for converted users and over 250,000 for not converted, so it is indeed large enough and satisfies said chi-square test assumption. The large sample ensures the random sampling of data. The variables that we’re focusing on within our data are the new and old advertisement pages as well as the conversion rate, therefore we assumed that our data did not experience the flicker effect; which would cause correlation because the consumer would view both the new and old ad page. During our data analysis, we removed all repeated observations thus observations are independent and don’t hold any correlation.

***Evaluations***

As noted in the data visualization and exploratory data analysis section, we conducted preliminary analysis such as creating said table of means and a five number summary. Based on these analyses, we see that the standard deviation of the old page’s conversion rate is 0.3254138 and for the new page’s conversion rate is 0.3235636, as seen in Figure 6. Although we mainly used the conversion rate variable, we firstly implemented basic analytical methods to assure that our conclusions were valid.

The chi-square test tested the null hypothesis that states that the old and new ad page would not result in a significant difference in the conversion rates (formula found in Figure 8). On the other hand, the alternative hypothesis states that the new advertisement and old advertisement would result in a significant difference in conversion rates. After running the chi-square analysis, we see that the p-value was 0.19, which is significantly higher than our alpha value of 0.05. As the p-value is much greater than the significance level, we fail to reject the null hypothesis and further concluded that old advertisement and new advertisement do not have a substantial difference in the conversion rate. There was not sufficient evidence to conclude that the conversion rates between the old and new advertisement differentiate to the point where the company should use the new advertisement. From a business standpoint, using the new advertisement would not necessarily attract more website visitors, moreover not increasing profits or lead to higher demands of what is being advertised in the ads. Also, an interesting observation to note is that despite the old ad page holding more conversions than the new one, both old and new ad pages held higher numbers for web pages that did not result in conversion compared to the ones that did (Figure 5).

To further analyze the difference of conversion rates between the old ad page and new page we did a 95 % confidence interval The 95% confidence interval estimates said difference between the two webpage conditions and allows us to see whether the chi-square test observations were accurately concluded. The 95% confidence interval is from -0.0007144889 and 0.0039880781, as shown in figure 7. The 95% of the confidence intervals contains the true difference of 0. The result of the confidence interval further confirms our chi-square test result that old ad page and the new ad page have no significant difference in conversion rate.

The results of these analyses indicate that the old ad page design slightly appeals more to the website visitors as the difference between the old and new web page conversion is actually minimal. It seems that the website design can simplify the composition, while the new website design can be slightly more distracting to the visitors. The investment in developing a new ad design as is currently depicted would be ineffective and purposeless.

***Recommendations***

A recommendation for a different statistical technique would be utilising machine learning in order to perform the test with a larger random sample. Since the A/B test resulted in no difference between the old and new page, and time has already been spent on the comparison of these two ads, we would recommend oversampling the data that has already been collected. With a larger sample which would make up for the lost time and the use of a method such as kNN, you can predict if a user will convert or not and run the A/B test again to see if updating the advertisement will be worth it in the long run.

Another future statistical technique that might be appropriate to try would be a different method of A/B Testing described in the Background section, A/B/N or Multivariate Testing. This method of testing would allow for the company to test a variety of different versions of the advertisement with various alterations to find the features that generate the highest conversion rates. Due to our conclusion, we’d recommend our client to make more drastic changes to the website since there was not much variance with the conversion rates comparing the old and the new website design. The changes should stem from research in consumer experience and purchase patterns to create a page that holds higher conversion rates in general. We would also recommend the company to analyze the demographic of the users and see what age group, sexual orientation or ethnicity they fall into in order to uniquely alter the ad for the target audience.

To ensure accurate results of the A/B test, the company should run the test as long as they can afford to do so. The longer collection periods will ensure accurate overall results that account for spikes due to the day of week, time of day and the season. This is also generally a good idea so that the company has a baseline to compare with when they perform any future A/B Tests. People might have different preferences for different seasons of the year.

**Works Cited**

Brebion, Anthony. “How to Avoid Flickering (Flash of Original Content) in A/B Tests.” *AB Tasty*, 9 Oct. 2018, [www.abtasty.com/blog/ab-testing-flicker-effect/](http://www.abtasty.com/blog/ab-testing-flicker-effect/).

Cabot, Jordi. “A/B Testing for Dummies.” *Nelio Software*, 9 Apr. 2018, neliosoftware.com/blog/ab-testing-dummies/.

Chapman, Cameron. “A Beginner's Guide To AB Testing: An Introduction.” *Neil Patel*, 15 Feb. 2019, neilpatel.com/blog/ab-testing-introduction/.

Deshpande, Bala. “2 Key Assumptions to Be Aware of before Applying the Chi-Square Test.” *Analytics Made Accessible: for Small and Medium Business and Beyond*, Soliton Technologies, [www.simafore.com/blog/bid/56480/2-key-assumptions-to-be-aware-of-before-applying-the-chi-square-test](http://www.simafore.com/blog/bid/56480/2-key-assumptions-to-be-aware-of-before-applying-the-chi-square-test).

Parmar, Divya. “Introduction to A/B Testing.” *Data Science Central*, DSC, 23 May 2017, [www.datasciencecentral.com/profiles/blogs/introduction-to-a-b-testing](http://www.datasciencecentral.com/profiles/blogs/introduction-to-a-b-testing).

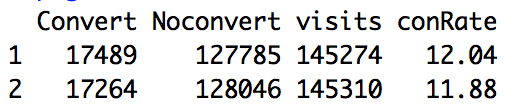
Piccinini, Page, director. *Assumptions and Types of A/B Testing*. *R*, *2018* campus.datacamp.com/courses/ab-testing-in-r/chapter-3-experimental-design-in-ab-testing?ex=5.

Shukairy, Ayat. “Momentum Behavior: Are Visitors in Love with Your Old Site Design?” *Invespcro*, Invesp, 2017, [www.invespcro.com/blog/momentum-behavior-are-visitors-in-love-with-your-old-site-design/](http://www.invespcro.com/blog/momentum-behavior-are-visitors-in-love-with-your-old-site-design/)

Zhang, Luyuan. “A/B Testing.” *Kaggle*, Kaggle Inc, 25 June 2018, [www.kaggle.com/zhangluyuan/ab-testing](http://www.kaggle.com/zhangluyuan/ab-testing).

***Appendix***

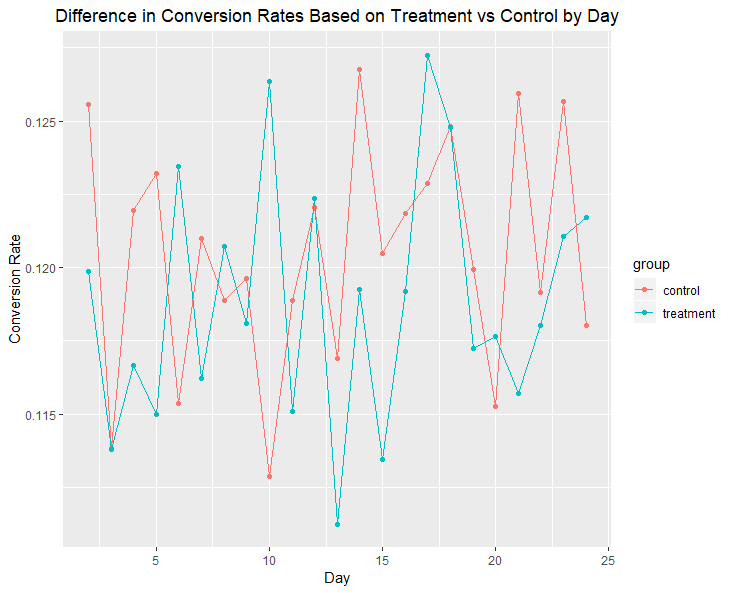
**Figure 1.** This shows the number of converted and unconverted values, the total count and rate for the old, denoted as 1, and new page, denoted with a 2.

****

**Figure 2.** The average conversion rate between old and new pages

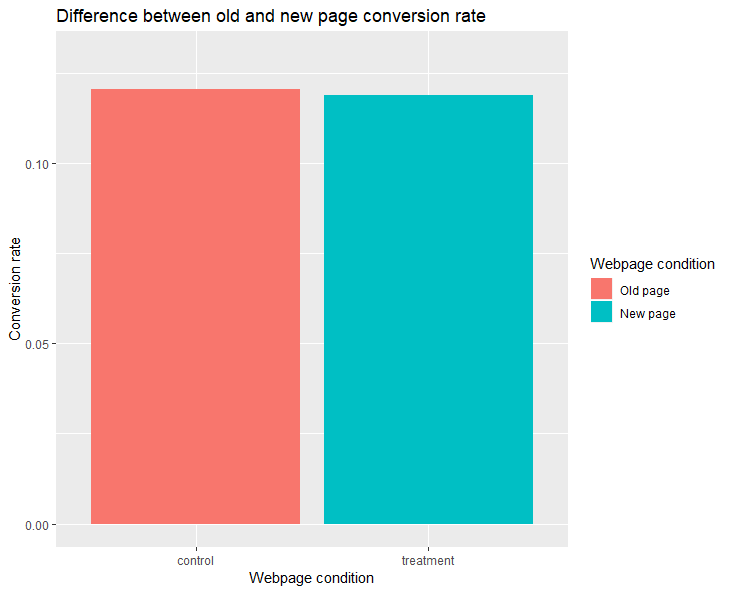
****

**Figure 3:** Conversion rate over time

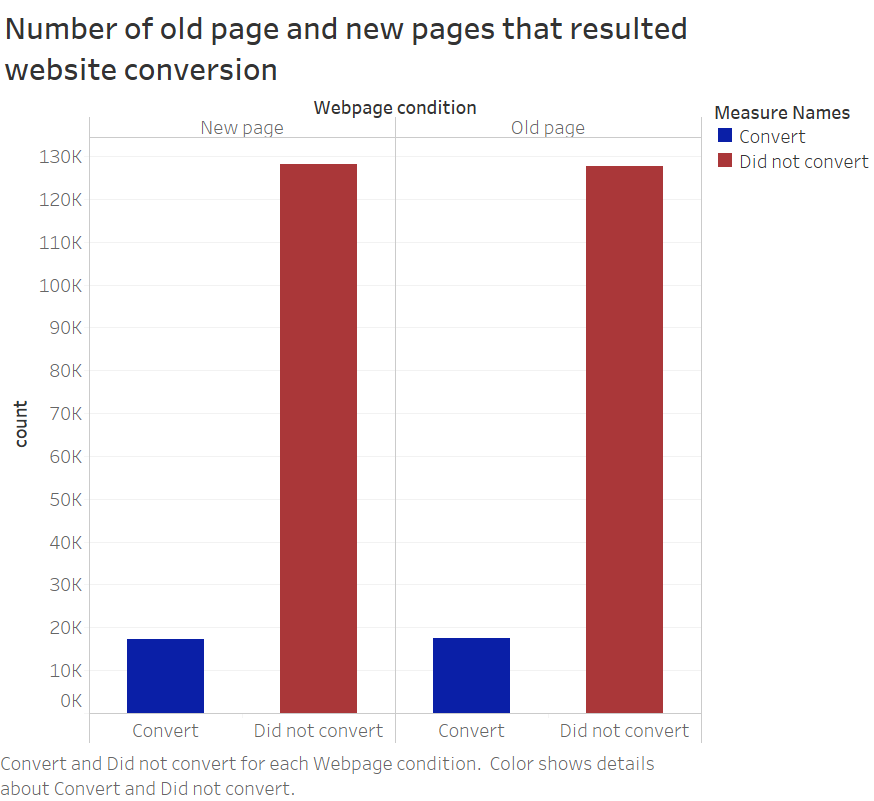


**Figure 4: Bar plot of conversion rate**

This figure shows the difference in conversion rates between the old and the new page and you can see that they are almost equivalent in length.

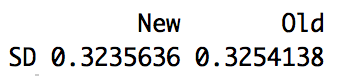


**Figure 5:** Count of web pages that resulted website conversion and didn’t result website conversion



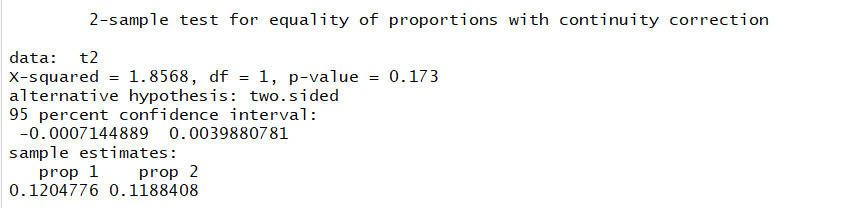
**Figure 6:**

The standard deviation of conversion rates between old and new pages

****

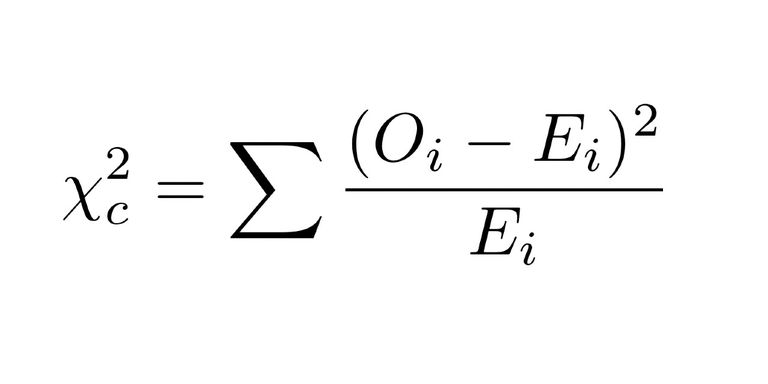
**Figure 7**

Confidence interval of the Chi Squared test.

****

**Figure 8**

Formula of the Chi-square test

****

**Table 1**

Summary statistics table of our data frame with the 5 number summary, mean, standard deviation, sample size and missing values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | User ID | Timestamp | Group | Landing Page | Converted |
| 5 Number Summary | Min:  630000  1st Q: 709035  Median: 787996  3rd Q: 866956  Max:  945999 | Min: 1/2/2017  1st Q:  1/8/2017  Median: 1/13/2017  3rd Q: 1/19/2017  Max: 1/24/2017 | n/a | n/a | Min: 0.0  1st Q: 0.0  Median: 0.0  3rd Q: 0.0  Max: 1.00 |
| Mean | 788005 | 13.07 (date) | n/a | n/a | 0.1196 |
| SD | n/a | 6.363515 (date) | n/a | n/a | 0.3244903 |
| Sample Size | 290,584 | 290,584 | Control: 145,274  Treatment:  145,310 | Old page: 145,274  New page: 145,310 | 0: 255,831  1: 34,753 |
| Number of Missing Observations | 0 | 0 | 0 | 0 | 0 |