# Pipeline for GOOD case study

Lingyi Tan
Mar.19th 2019

## Method

Firstly, data cleaning. I found that there are some disqualified answer in both test and control group, so my first step is to remove those disqualified data. Meanwhile I checked the distribution of disqualified data among two groups. Control group has significantly more disqualified data than test group, which means respondents exposed to the foundation's content were more willing to conduct the survey.

Next, data exploration. Since all the features are discrete variables, I firstly generated a pivot table to describe the complete data(which is the cleaned data). Result see the first sheet “Summary”. According to this result, I then calculated the statistics about the top 2 box in question 1 to 3, and used 2-sample test to see if the percentage of top 2 box were different in test and control group. Here we are comparing the crude results, Result see the second sheet “Crude”, which is the summary of top 2 box ratings that are not classified by any features.

**Let’s now use the word “intervention” to represent the expose of the foundation's content.** When comparing the top 2 box rating rates for Q1, I found that the difference between test and control was statistically significant at the 90% CI as the p-value for 2-sample test for equality is much less than 0.1.See R code comments and result. While the difference of top 2 box rating rates between test and control group for Q2 and Q3 are not significant, as both of their p-value are greater than 0.1. See R code comments and result.

To better present the result to audience who is not a data analysis expert, I used R to do the data visualization. See graph: Q1Q2Q3.png. It is very obvious that the percentage of top 2 box ratings for Q1 is significantly reduced in test group compared with control group. Based on above result, we can claim that in generally, the intervention is effectively increasing people’s awareness of importance of science, especially for those who didn’t think science was important. However, the intervention doesn’t help a lot in helping people realize the importance of funding of science experiences neither the statement that "Making science more accessible and inclusive is essential to a functioning society." Though there was no sign of the intervention effect on Q2 regarding to the top 2 box ratings, we could still find that the intervention had a great impact on the percentage of respondents who thought funding of science experiences was very important or absolutely essential.

Considering of the potential confounding, I then stratified the data by features(gender,

age, Q4 and Q5) to see if those features were confounders in the analysis. I firstly did the data visualization to see the stratified result regarding to Q1,Q2 and Q3. When stratified by gender, the impact of intervention for male and female seemed different regarding to all three questions. See graph: Gender.png. I then generated the stratified table for gender See the third sheets “Stratified_Gender” and calculated the relative risk of top 2 box ratings for male and female respectively. It’s obviously that the effect of intervention on Q1 was much stronger in male than female, and the effect for Q3 was even opposite in male and female. Similar manipulation were done to age, Q4 and Q5. See graph: Age.png, Q4.png and Q5.png For age feature, I classified data by age<= 29 and age>30 based on the data visualization that age group (18-29) had a different pattern compared with other groups. We can find that people older than 30 were more likely to be influenced by the intervention regarding to Q1 and Q3 compared with people younger than 30. See sheet “Stratified_Age” When stratified by Q4, there is not significant difference regarding to Q1-Q3 among people with different answers. When stratified by Q5, the pattern of answer to Q3 for respondents who didn’t read scientific content at all weren’t changed a lot by intervention, while the others were. So I divided data into respondents who didn’t read scientific content and the others, and checked their relative risks. Again, the effect was opposite in two groups. See sheet “Stratified_Q5” The intervention worked well for people who read scientific content content, but not for people who didn’t read scientific content at all.

## Conclusion

1. Generally, the intervention(exposed to the foundation's content) increased people’s awareness of importance of science, but helped little in evoking people’s realization of the importance of funding of science experiences and the agreement of statement that "Making science more accessible and inclusive is essential to a functioning society."
2. The effect of the intervention differed in different subgroups. For example, it made a greater positive impact on male compared with female(for both Q1 and Q3), and a greater positive impact on people older than 30(for both Q1 and Q3) and a greater positive impact on respondents who read scientific content at least sometimes(Q3).
3. For a brand who is supposed to improve science education/funding, the foundation did a great job at evoking people’s awareness of importance of science. They might want to focus more on the advertising of education/funding as this is their principle part.
4. If the foundation has its products or training, then the target customers/potential customers for the products would be mainly male, people older than 30 years, and people who read scientist content.