Samba TV Technical Exercise

Andrew Candela

November 3, 2016

Abstract

In this presentation we discuss the issues with XYZ's interpretation of a test of the efficacy of local translations of their web content. I have found that their interpretation suffers from a mischaracterization stemming from an effect similar to Simpson's Paradox. While the results of XYZ's analysis are not necessarily reversed when looking at subsets of the population, there are anomalies in the distribution of test and control users within certain cohorts that distort the analysis when taken in aggregate. We will also consider an algorithm that can alert future consumers of this type of analyses when a mistake of this kind is a possibility. Finally, we will discuss how to quantify the strength of the observed effect of this test. We will then use that to determine if XYZ should in fact implement local translations of their website, rather than one Spanish language translation.

Description of the Problem

Company XYZ is trying to decide if a local translation of their web content has a positive effect on conversion rates of users. They conduct an experiment meant to help estimate these effects. The overall results are found in Table 1 below. Based on these results it seems that XYZ has concluded that the test (test = 1) population performs worse than the control (test = 0) population. Let's take a closer look by examining Figure 1 below.

It looks like the Response Rate (RR) for test is slightly higher than RR for control in some countries. Notice that the response rate for Spain is much higher than that of any other country, and Argentina and Uruguay have much lower response rates than those of the other countries. Now notice that Argentina and Uruguay have many more test users than control users, and that Spain has no test users.

Since the two countries with the worst response rates (overall) contribute far more (relatively) test users than control users, the average RR for the test population is dragged down. Add to this the fact that Spain contributes a relatively large number of control users and no test users, and you end up with a recipe for misinterpretation when you aggregate

test ($\boldsymbol{T}$)	conversions	users	response rate (RR)
0	13077	237093	0.055156
1	9367	215774	0.043411

Table 1: Overall Test Results - RR for test is lower overall than RR for control

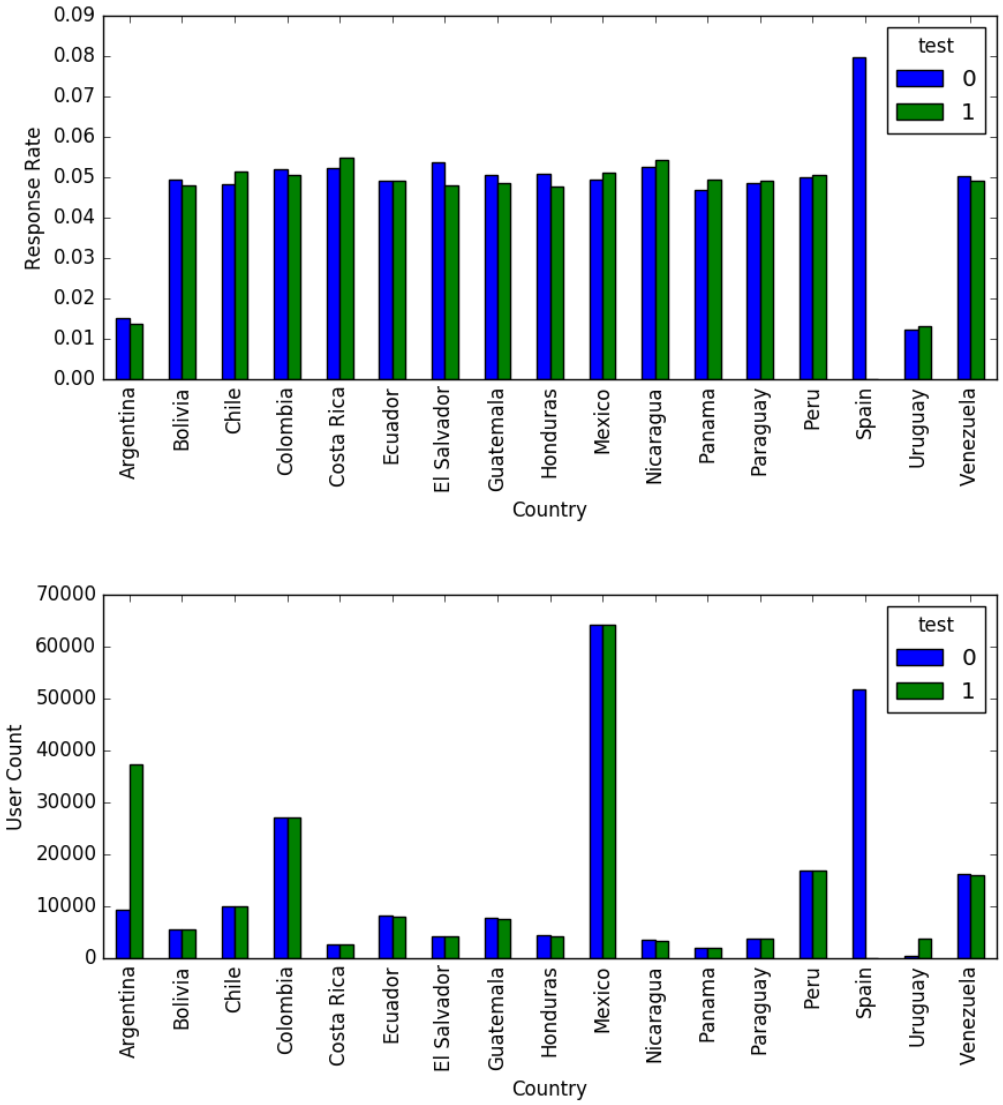

Figure 1: User Counts and Response Rates by Country - Some countries have a much higher proportion of test users than control users.

RR over country. This is the mistake I believe XYZ made when they interpreted the results of this experiment.

Solution and Preventative Measures

Since the problem stems from an uneven distribution of test and control users across countries that drive a large portion of the variation of the target variable, we would want to do one of the following:

- Treat each country as a separate experiment
- Throw countries with relatively uneven amounts of test and control users out of the experiment

I prefer to use the first approach. Since each country has its own translation, we can look at the results of that country independently of the others, and decide for each one if there is a difference between the test and control groups.

An Algorithm to Detect Future Issues

We can develop an algorithm to detect uneven distributions of test and control users among each level of all of the explanatory variables in this experiment. Let's pretend for now that we are given a pandas dataframe with the experiment results, and an array of the explanatory variable names.

```
from __future__ import division

def check_user_distribution(dataframe,explanatory_variables):
  flag=0
  #create a convenience row to help me sum up control users
  dataframe['control']=[1 if i==0 else 0 for i in dataframe['test']]
  for v in explanatory_variables:
    comp_df=dataframe.groupby(v).sum()[['test','control']]
    for i in comp_df.iterrows():
      tc_ratio=i[1]['test']/i[1]['control']
      if (tc_ratio > 1.10) | (tc_ratio < .9):
        print ('Watch out! Test/control ratio for {} is: {}'.format(i[0],tc_ratio))
        flag=1
  if flag=1:
    return False
  else:
    return True
```

This function would iterate through each level of each explanatory variable and print a warning if any level has an uneven number of test and control users. I chose the limits of 1.1 and .9 arbitrarily here. While I am confident that there is a more rigorous way to classify this type of issue programmatically, this algorithm has the advantage of being very easy to implement and understand. I'd love to hear how other folks have approached this problem!

Should XYZ Create Local Translations?

In order to decide if XYZ should create local translations of its content, we must first decide how strong we believe the effect that $\boldsymbol{T}$ has on $P(\boldsymbol{C} = 1)$. I propose running a very simple logistic regression for each country individually. We can use the 95% confidence interval computed to learn how volatile we would expect future response rates to be.

The Model

I will use the following model for the probability p that a user converts:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 test$$

Fitting this model to the data from Mexico with the statsmodels python package gives the following results:

```
                           Logit Regression Results
==============================================================================
Dep. Variable:             conversion   No. Observations:               128484
Model:                          Logit   Df Residuals:                   128482
Method:                           MLE   Df Model:                            1
Date:                Tue, 01 Nov 2016   Pseudo R-squ.:               3.751e-05
Time:                        19:14:23   Log-Likelihood:                -25634.
converged:                       True   LL-Null:                       -25635.
                                        LLR p-value:                    0.1655
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
test           0.0354      0.026      1.387      0.166        -0.015     0.085
intercept     -2.9551      0.018   -162.417      0.000        -2.991    -2.919
==============================================================================
```

We can see that while the coefficient for the test variable is positive, the 95% confidence interval for that coefficient contains 0. Also, we can see the p-value of this test is 0.1655. While I would say that $test = 1$ is likely to have a positive impact on probability of conversion, I can't say it with great confidence. Depending on XYZ's appetite for risk, and the value of an additional likelihood of conversion, XYZ may decide to implement local translations for some countries.

Is it Worth it?

In order to decide if XYZ should create localized translations, we must first agree upon the following:

1. Cost - How much it will cost to produce additional translations
2. R_i - How much additional revenue we expect by providing the localized translations to the ith country

I propose the following cost model:

$$Cost = F + \sum_i T_i$$

where T_i is the cost of supporting the ith country's translation, and F is a fixed cost of implementing the translations.

We can estimate R_i by examining the results of XYZ's experiment. For example, we estimate that in Mexico the probability of conversion for a control user is 0.049 and the probability of conversion of a test user is 0.051. We define $\lambda_{Mex} = 0.002$ as the difference in response rate of the test group and control group. If the average Mexican conversion generates δ_{Mex} revenue, we would expect to see an additional revenue per user of $\lambda_{Mex}\delta_{Mex}$ by switching Mexico to the localized translation. If the expected number of distinct users per time period for Mexico is given by D_{Mex}, then we can conclude that $R_{Mex} = \lambda_{Mex}\delta_{Mex}D_{Mex}$.

Now we can repeat this experiment for each country c in Θ: the set of all countries that have a higher RR for test than control. Then we can see if the following holds:

$$F + \sum_{c\in\Theta} T_c < \sum_{c\in\Theta} \lambda_c\delta_c D_c \tag{1}$$

If the above inequality holds, then XYZ should consider implementing the localized translations for the countries in Θ.

Conclusion

Company XYZ has run an A/B test to determine if users respond better to web content translated by a local rather than a Spaniard. The test was incorrectly interpreted as a failure, as the overall response rate of the test group was lower than that of the control group. The interpretation of the results taken in aggregate were influenced heavily by an uneven distribution of test and control users in countries with abnormally high or low overall response rates. Even when blocking for this and examining the results by country, strong evidence that the test group outperforms the control group was not observed. Despite a lack of strong evidence that the test group performed better than the control group, we may still wish to decide if it is worth it to implement individual translations. The methods we discussed would give an effective criteria to use to make that decision.

We also discussed an algorithm that will check for uneven distributions of test and control users among all levels of the explanatory variables and alert the user if an inequality exists.