Thumbtack Programming Challenge
Andrew J Smith ("Drew")
October 29, 2013

Thumbtack Split Test Analysis

In modern web application frameworks understanding your users and how they interact with your website is of key importance. A well designed User Interface (UI) can lead to a great User Experience (UX) and more business for your company. One method for testing new and 'improved' versions of a UI is through A/B Split Testing.

In most cases a series of significant changes are made to a baseline webpage or series of webpages and displayed to different cohorts of users at random. Each case is tracked separately and after the trial period is over the statistics are compiled and analyzed for significance.

The Thumbtack data set provided in the challenge includes a baseline page for the provider quote form and 4 variations (Table 1). Let's dive in a see if the conclusion that Variation 3 results in more providers submitting quotes is valid.

Bucket	Quotes	Views
Baseline	32	595
Variation 1	30	599
Variation 2	18	622
Variation 3	53	606
Variation 4	38	578

Table 1. Thumbtack Split Test Dataset including the Baseline and Four Variations

At first glance the numbers look good for Variation 3 (see Appendix A). The Chi-squared $X^2(1, n=1,194) = 5.19$, $p \leq 0.023$ is above the critical limit of 3.841 for a $p<0.05$ showing that it is statistically significant at the 95% level. Variation 1 and Variation 4 show no signs of significant change from the baseline but Variation 2 on the other hand shows a significant $X^2(1, n=1,217) = 3.92$, $p \leq 0.048$. Unfortunately the direction of the change is negative meaning that the users did not react positively to the changes on the quote form in Variation 2.

The goal of this Split Test was to encourage more providers to submit quotes but there are many pit falls when it comes to A/B split testing that can lead a company to believe that the statistically significant split was successful. Let's pose a few questions that can help reveal whether we accept the test as valid or to declare an invalid split test.

- Were all of the providers in the study selected at random?
- Were the variations chosen in random order or cycled through in sequence (V1, V2…)?
- Was the variation quote form only provided to those providers who reached the quote form stage of the process or was the variation cohort assignment selected at login?
- Were the variations displayed at the same time of day/day of week/month?
- Which categories of providers were included in the variation cohorts? Are they spread evenly across all categories of providers?
- Did any provider categories prefer a specific variation more than Variation 3?
- Did Thumbtack run a promo or marketing exercise recently that could skew the results?
- Were there any regional variations in preference? SF vs. NY?
- Where there any significant age or gender preferences between the tested quote forms?

While it can be difficult to get a truly random split test, careful planning and multiple test cycles can give clearer answers to the questions that drive business through a web site.

Thank you reading my analysis of the Thumbtack Split Test Programming Challenge and I look forward to receiving feedback from Thumbtack on my code and this analysis.

Appendix A: Results of ThumbTackSplitTest.py

Starting Thumbtack Split Test Analysis on: Tue Oct 29 12:58:34 2013

ThumbTackSplitTest Version: 2013.10.29 demo
NumPy Version: 1.7.1
SciPy Version: 0.12.0

Results of the A/B Split Test

Data:
[[32. 595.]
 [30. 599.]
 [18. 622.]
 [53. 606.]
 [38. 578.]]

Variation #1
Observed: [32. 30.]
Expected: [31. 31.]
Trials: [595. 599.]
ChiSq = 0.0645161290323
p-value = 0.79949536187
This test is not statistically significant

Variation #2
Observed: [32. 18.]
Expected: [25. 25.]
Trials: [595. 622.]
ChiSq = 3.92
p-value = 0.0477148802374
This test is statistically significant at the 95% confidence level

Variation #3
Observed: [32. 53.]
Expected: [42.5 42.5]
Trials: [595. 606.]
ChiSq = 5.18823529412
p-value = 0.0227402961153
This test is statistically significant at the 95% confidence level

Variation #4
Observed: [32. 38.]
Expected: [35. 35.]
Trials: [595. 578.]
ChiSq = 0.514285714286
p-value = 0.47328946538
This test is not statistically significant