Anubhav Rana: Galvanize technical exercise Q3

Baseline: 32 quotes out of 595 viewers
Variation 1: 30 quotes out of 599 viewers
Variation 2: 18 quotes out of 622 viewers
Variation 3: 51 quotes out of 606 viewers
Variation 4: 38 quotes out of 578 viewers

Bucket	Number of Quotes	Total number of viewers	Conversion Rate	Conversion Rate %
Baseline	32	595	0.05378151261	5.378151261
Variation 1	30	599	0.05008347245	5.008347245
Variation 2	18	622	0.02893890675	2.893890675
Variation 3	51	606	0.08415841584	8.415841584
Variation 4	38	578	0.06574394464	6.574394464

Table 1: Table showing the conversion rates for the different variations of the quote form

Are the differences in number of quotes statistically significant?

To test the significance of the number of quotes the z - test can be used. The null hypothesis would be that the differences in quotes has occurred by chance and that the differences are not statistically significant. The alternate hypothesis would be that the differences did not occur by random chance and that the they (the differences) are statistically significant. Since the distribution of the test-statistic can be approximated by the normal distribution (central limit theorem) , the z-test is a valid test to check the statistical significance of the differences in quotes.

In order to test the hypothesis the test-statistic will need to be calculated. By then comparing the test-statistic to our significance level (commonly denoted by α), we can deduce whether the null hypothesis can be rejected or not.

If the test-statistic is greater than the z-score, that tells us that the null hypothesis can be rejected and that our results are statistically significant. A good value for α is 0.05 or 5%. For this significance level the value of the z-score (critical value) is 1.645. The z-score for the significance level of 0.01 or 1% is 2.326.

The equation used to calculate the test statistic is:

$$z = \frac{p - p_c}{\sqrt{\frac{p(1-p)}{N} + \frac{p_c(1-p_c)}{N_c}}}$$

Where, p is the conversion rate of the baseline and p_c is the conversion rate of the variation. N is the number of viewers in the sample of the baseline and N_C is the number of viewers in the variation. The python code for calculating the test-statistic and p-value can be found below.

```
import numpy as np
from scipy import stats

#Creating list with quotes and Viewers
N = [595, 599, 622, 606, 578] # List of total Viewers for each variation
Q = [32.0, 30.0, 18.0, 51.0, 38.0] # List of Quotes
m = float(Q[0]/N[0]) #The baseline conversion rate

#initializing test statistic and conversion rate lists
CR = []
t_stat = []

#Calculating conversion rate
for i in range(1, len(Q)):
        CR.append(float(Q[i]/N[i]))

#Calculating test statistic
for i in range(len(CR)):
        t_stat.append((CR[i]-m)/np.sqrt((CR[i]*(1-CR[i])/N[i])+(m*(1-m)/N[0])))
```

Bucket	Conversion Rate %	Test-statistic
Baseline	5.378151261	N/A
Variation 1	5.008347245	-0.2874689689
Variation 2	2.893890675	-2.158662085
Variation 3	8.415841584	2.098986601
Variation 4	6.574394464	0.8750517255

Table 2: Table showing the conversion rate and test-statistic for the different variations of the quote form.

From table 2 above, we can see that the null hypothesis can only be rejected for variation 3, where 2.099 > 1.645. Coincidentally, the conversion rate for this variation is also the highest. This tells us that we can say with 95% confidence that the differences in quotes for variation 3 were not due to random chance. The same cannot be said for the other variations and the null hypothesis cannot be rejected, i.e, the differences in quotes for the other variations could be due to random chance.

Questions to ask:

Is the sampling random? Were the viewers for the different variations taken from the same sample? For example, if one variation had mostly women and the other had mostly men, it would be difficult to compare the two versions, since one side may prefer certain aspects more than others.

Were the viewers all unique viewers? It is possible that a quote was sent out but that customer viewed the page several times. This would affect the conversion rate as that person's view would be counted multiple times.

One more question regarding viewers would be about the location. Were the different variations shown to all the same locations? The traffic to the forms for different locations would affect the conversion rates and certain locations may yield higher conversion rates in general.

Were the invites evenly spread through the different categories? It is possible that certain categories have a higher conversion rate and this could also affect the conversion rates for the different variations. The category can also be cross examined with the location, since it's possible that certain categories are higher in demand at certain locations.

Did a particular viewer only see one variation of the page? If one viewer sent a quote but saw different variations of the page, then that would invalidate the results as we he/she could have sent the quote after viewing a certain variation which changed on subsequent visits.

Another question would be about the specifics in the variations. It is possible that the variations resulted in the customers sending a quote, just not through the quote form provided on the website.

One improvement to the experiment design could be to add more samples (viewers) for the different variations. This would result in better testing and could further ensure that the sampling was random.