Split Test Analysis, Question 3

After reading the problem a few times, it looks like we can organize the information with the following:

1. The baseline can be interpreted as a result we typically expect, if we hadn't performed any variations to the way the information flow is carried out between the customer and provider. In other words, the baseline is the results we obtain, on average, if no manipulations to the process were made.
2. The quotes are meaningless in absolute terms. Let's put the quotes in relative terms, comparing them with the total number of viewers. Therefore, let's generate a new metric, called 'Success Ratio' (SR) that represents the ratio of quotes to viewers.
3. The statistical question that deems appropriate is: are the variations drastically affecting results? In other words, are their success ratios statistically different from what we normally expect, AKA the baseline?

So, let's first get started by extracting the data from the baseline and variations, cleaning them up as needed.

In [2]:

```
import csv
path = 'data.csv'

x = []
y = []

with open(path, newline = '') as f:
    reader = csv.reader(f)
    for row in reader:
        x.append(row[1])
        y.append(row[2])
```

In [3]:

```
x
```

Out[3]:

```
['Quotes', '32', '30', '18', '51', '38']
```

In [4]:

```
y
```

Out[4]:

```
['Views', '595', '599', '622', '606', '578']
```

In [5]:

```
baseline = [(int(x[1]), int(y[1]))]
```

In [6]:

```
baseline
```

Out[6]:

```
[(32, 595)]
```

In [7]:

```
variations = [(int(x[2]), int(y[2])),(int(x[3]), int(y[3])),(int(x[4]),int(y[4])), (int(x[5]), int(y
[5]))]
```

In [8]:

```
variations
```

Out[8]:

```
[(30, 599), (18, 622), (51, 606), (38, 578)]
```

Now let's define the success ratio:

In [9]:

```
def success_ratio(input):
 sr = []
 for i, v in input:
  sr.append(float(i/v))
 return sr
```

In [10]:

```
expected_value = success_ratio(baseline)
```

In [11]:

```
expected_value
```

Out[11]:

```
[0.05378151260504202]
```

To make sure we all understand what the SR is saying, it can be interpreted as the following: on average, around 5.4% of provider actively respond to notifications, assuming we did not run any manipulations to the way the information flow is carried out.

In [13]:

```
variation_set = success_ratio(variations)
variation_set
```

Out[13]:

```
[0.05008347245409015,
 0.028938906752411574,
 0.08415841584158416,
 0.0657439446366782]
```

Here, we have the SR's of the four manipulations that were carried out. The question to ask now is: are these SR's statistically significant? In other words, are they purposefully generating desired results, or are they just tiny variations from the expected result?

If we assume that the variations were all conducted independently, in isolated events, and if we also assume that the baseline is generated from a normal distribution, then we can run a one-sample t-test here.

A One Sample T-test allows us to work with small sample sizes to determine whether the manipulations that were made in the variation set are, on average, statistically significant from the expected value.

So, let's break the problem down into two hypotheses:

Null hypothesis: Variations do not impact the SR Alternate hypothesis: Variations significantly impact the SR

Depending on the t-statistic that we obtain from the data, we can figure out whether to reject or not reject the null hypothesis. If we can reject the null hypothesis, then perhaps the variations are purposefully giving us something we want.

In [17]:

```
import scipy.stats

onesample_results = scipy.stats.ttest_1samp(variation_set, expected_value[0])
onesample_results
```

Out[17]:

```
Ttest_1sampResult(statistic=0.2942705909906809, pvalue=0.7877376338259627)
```

Our t-statistic is ~0.294, and pvalue is ~0.788.

The p-value is quite high, and definitely something to look at. If we average out the variation set, we get:

In [20]:

```
float(sum(variation_set)/len(variation_set))
```

Out[20]:

```
0.05723118492119102
```

Conclusion

What the p-value is saying is that 79 out of 100 times we can expect to see an average of SR's to be greater than or equal to 5.7%, the average SR from the variation set.

If almost 80% of expectations are above our variation set, is our variation set anything special then? Well, let's look at the t-statistic. Our t-score is 0.294. In order to interpret this result, we have to calculate the degrees of freedom of our sample sets: 1 + 4 - 2 = 3. Let's also set our alpha, which is a metric that determines how confident we are, to 0.10. We want a one-tailed t-test, since we are only concerned with moving the SR in one direction, i.e. we want to increase it.

If you select the row where df = 3, and select the column where one-tail = .10, then you'll find that the t-test = 1.638, which is the threshold you must pass in order to deem an experiment as statistically significant. Since our t-score is 0.294, which is less than 1.638, we can conclude that we can't reject the null hypthosis. In other words, our variations didn't generate results important enough to claim that they made significant results.

Points/Questions to consider:

Just because we obtained results that weren't desirable does not mean we should stop making experimenting. Few reasons why:

1. The size from both samples is too low. If we had more data, we would be able to generate results that would be much richer and conclusive than what we have now.
2. We are making the assumption that the baseline was generated as the average result. If this assumption is wrong, then our t-test would be invalid, as it won't accurately depict how off-set the variation set is from the mean.
3. I'm curious what the methodology was under variation 3, as that generated a relatively high SR (8%).