P1_StroopEffect

December 22, 2017

1 Project 1: Stroop Effect

Prepared By: Samba Reyes Njie Jr.
When: December 2017

1.1 Background Information

In a Stroop task, participants are presented with a list of words, with each word displayed in a color of ink. The participant's task is to say out loud the color of the ink in which the word is printed. The task has two conditions: a congruent words condition, and an incongruent words condition.

- In the congruent words condition, the words being displayed are color words whose names match the colors in which they are printed.
- In the incongruent words condition, the words displayed are color words whose names do not match the colors in which they are printed.

In each case, we measure the time it takes to name the ink colors in equally-sized lists. Each participant will go through and record a time from each condition.

1.2 Questions For Investigation

1. What is our independent variable? What is our dependent variable?
2. What is an appropriate set of hypotheses for this task? What kind of statistical test do you expect to perform? Justify your choices.

Go to this link, which has a Java-based applet for performing the Stroop task. Record the times that you received on the task (you do not need to submit your times to the site.) Now, download this dataset which contains results from a number of participants in the task. Each row of the dataset contains the performance for one participant, with the first number their results on the congruent task and the second number their performance on the incongruent task.

3. Report some descriptive statistics regarding this dataset. Include at least one measure of central tendency and at least one measure of variability.
4. Provide one or two visualizations that show the distribution of the sample data. Write one or two sentences noting what you observe about the plot or plots.

5. Now, perform the statistical test and report your results. What is your confidence level and your critical statistic value? Do you reject the null hypothesis or fail to reject it? Come to a conclusion in terms of the experiment task. Did the results match up with your expectations?

6. Optional: What do you think is responsible for the effects observed? Can you think of an alternative or similar task that would result in a similar effect? Some research about the problem will be helpful for thinking about these two questions!

1.3 Prepwork

```
In [4]: # Import Libraries
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        from scipy import stats
        %matplotlib inline
```

```
In [5]: # Read in data
        data = pd.read_csv('stroopdata.csv')
```

```
In [41]: ##### print("Table 1: Data on Stroop Effect Participants from Both Treatments")
         data
```

Out[41]:

	Congruent	Incongruent	Difference
0	12.079	19.278	7.199
1	16.791	18.741	1.950
2	9.564	21.214	11.650
3	8.630	15.687	7.057
4	14.669	22.803	8.134
5	12.238	20.878	8.640
6	14.692	24.572	9.880
7	8.987	17.394	8.407
8	9.401	20.762	11.361
9	14.480	26.282	11.802
10	22.328	24.524	2.196
11	15.298	18.644	3.346
12	15.073	17.510	2.437
13	16.929	20.330	3.401
14	18.200	35.255	17.055
15	12.130	22.158	10.028
16	18.495	25.139	6.644
17	10.639	20.429	9.790
18	11.344	17.425	6.081
19	12.369	34.288	21.919
20	12.944	23.894	10.950
21	14.233	17.960	3.727
22	19.710	22.058	2.348
23	16.004	21.157	5.153

1.4 1. Identifying Variables and Problem Formalization

Independent Variable: Congruent Word Performance Treatment, Incongruent Word Performance Treatment

Dependent Variable: Difference in word performance

One of the phenomena we can test is if there is an associative difference between reading color names and reading colors. As such, if we design a test to record completion times of participants for a series of words describing the colors they are assigned to, we will be testing the association of words to color. If we want to go beyond describing a relationship and studying causality, then we want to use *causal inference* to distinguish if disassociating words from color as a treatment group would provide any differences.

To make the problem well-posed, we list the following definitions:

- **Stroop Effect:** is a phenomenon demonstrating the interference or inhibition of a specific task. In particular, it is the difference of reaction times between incongruent and congruent treatment cases. The difference in brain activity between these two conditions (i.e., incongruent minus congruent) could reveal brain systems involved in the attentionally mediated resolution of the conflict between the habitual response of reading words vs. the task demands of naming the color of the words[3],[5].

- **Law of associative inhibition:** "If a is already connected with b, then it is difficult to connect it with k, [because] b gets in the way."[5] The inhibition is the prevailing idea behind the interference phenomenon in the Stroop effect, where the automatic association of reading color names ("a") and naming colors ("b") is interfered by the incongruent case.

- **Congruent words condition:** The treatment state in which participants record reading duration for words where the noun and presentation color are congruent, or match[5]. For a participant/observation i, their congruent word performance is θ_i, and the set of them for all observations Θ.

- **Incongruent words condition:** The treatment state in which participants record reading duration for words where the noun and presentation color are incongruent[5]. For a participant/observation i, their congruent word performance is ϕ_i, and the set of them for all observations Φ.

- **Word Performance Difference:** Difference between congruent and inconguent words conditon performances, which operationalizes the effect of interference of word reading upon color naming:

$$d_i \doteq \phi_i - \theta_i$$

- **Sample Mean Difference:** The mean of word performance difference scores across all observations:

$$\bar{d} \doteq \frac{1}{n} \sum_{i=1}^{n} d_i$$

- **Sample Standard Deviation Difference:** Standard deviation of difference scores across all observations:

$$s_d \doteq \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}_i\right)^2}$$

Our data is collected via the online Stroop test[1], which collects congruent and incongruent word task completion times for *each* participant. As such, for any given observation i, where $i = 1, ..., n$, the net effect of interference is observed as d_i, as described above.

1.5 2a. Hypothesis Formulation

We formulate the hypothesis as follows:

$$H_0 \ : \ \mu_d = 0$$

$$H_1 \ : \ \mu_d > 0$$

Our goal is to understand if associative inhibition (or interference)[5] between reading words and naming colors exists. In other words, we want to know if there is significant evidence for a Stroop effect. To do this, we want to understand if there exists significant evidence for an interference, which is modeled by the difference in response times between mismatched nouns and colors (incongruent word task) and matching nouns and colors (congruent word task).

As such, the above mathematical formulation of hypotheses is a sufficient representation. For one, the null hypothesis assumes that there is no difference between population congruent word performance and population incongruent word performance, which cannot be specifically proven for now but can be retained or rejected depending on test results. The alternate hypothesis is a one-tailed hypothesis due to the assumption that incongruent word task times will almost always be greater than their congruent counterparts.

1.6 2b. Statistical Test Formulation

Since both θ_i and ϕ_i are dependent on one another for each observation, we choose a **paired sample t-test**. There are three main reasons for why we choose this test. Let us discuss the simpler reason first:

Reason 1: We have $n < 30$. As such, using a z-test approach will not work, and a t-test approach is ideal.

Reason 2: The assumptions for a dependent t-test are met. The assumptions are as follows: [6],[7]

1. *The dependent variable must be continuous (interval/ratio).*

   For $D \doteq \{d_i \ : d_i \ \in \mathbb{R}\}$, the set of word performance differences for all observations, is continuous since any value is in the real space, i.e., $D \subset \mathbb{R}$.

2. *The observations are independent of one another.*

   Since the data is collected from independently participating participants, then for any participant i, j, where $i \neq j$, θ_i is independent from θ_j and ϕ_i is independent from ϕ_j.

3. *The dependent variable should be approximately normally distributed.*

   This is true due to successful goodness of fit tests, which we will show later when doing goodness of fit tests in the Visualizations section. If we assume normality or any distribution,

we will need a parametric test, for which a paired 2-sample t-test would suffice. However, if there is no distribution, a non-parametric test is needed, for which the Wilcoxon Signed Rank Test will suffice.

4. *The dependent variable should not contain any outliers.*

   We will check this in our visualization later.

Reason 3: The paired design is an effective experimental technique to inspect the difference in associative interference between a congruent and incongruent batch. This is because θ_i and ϕ_i are dependent on i, which exhibits a within-subjects design. A within-subjects design is an experimental design where the same group participates in two treatments[8]. Since a dependent t-test is a type of within-subjects design, the statistical test befits the design type.

1.7 3. Descriptive Statistics

```
In [21]: # Transform original data to include a column of pairwise differences per observation
         data['Difference'] = data['Incongruent'] - data['Congruent']

         # user-defined function to calculate mode
         def mode(data, axis=0):
             """Returns the mode of a pandas data frame."""
             modes = stats.mode(data)
             tmp = modes.mode[0]
             output = pd.Series(tmp)
             output.index = data.columns
             return output

         # Create dataframe of statistics using Pandas and stats libraries
         means = data.mean(axis=0)
         stdevs = data.std(axis=0)
         variances = stdevs**2
         medians = data.median(axis = 0)
         q1 = data.quantile(q= 0.25, axis=0)
         q3 = data.quantile(q= 0.75, axis=0)
         modes = mode(data)
         lens = data.count(axis=0)
         desc = pd.concat([means.rename('Means'),
                           stdevs.rename('Std Dev'),
                           variances.rename('Variance'),
                           q1.rename('First Quartiles'),
                           medians.rename('Medians'),
                           q3.rename('Third Quartiles'),
                           modes.rename('Modes'),
                           lens.rename('n')], axis=1).T
         print("Table 2: Descriptive Statistics")
         desc
```

```
Table 2: Descriptive Statistics
```

Out[21]:

	Congruent	Incongruent	Difference
Means	14.051125	22.015917	7.964792
Std Dev	3.559358	4.797057	4.864827
Variance	12.669029	23.011757	23.666541
First Quartiles	11.895250	18.716750	3.645500
Medians	14.356500	21.017500	7.666500
Third Quartiles	16.200750	24.051500	10.258500
Modes	8.630000	15.687000	1.950000
n	24.000000	24.000000	24.000000

We chose the mean as a measure of centrality and the standard deviation to be the measure of variability. The above table summarizes descriptive statistics for each treatment set, Θ and Φ, and the statistics of the difference of samples, $D = \Phi - \Theta$. We then get the following statistics (with rounding):

$$\bar{d} = 7.964792$$

$$s_d = 4.864827$$

1.8 4. Visualizations

```
In [82]: def pivot(data):
             """Pivots original data frame with indices."""
             data = data.reset_index(level=0)
             output = pd.melt(data, id_vars= 'index',
                              var_name= 'Type',
                              value_name= 'Completion_Times')
             output.columns = ['Observations', 'Type', 'Completion_Times']
             return output

         data2 = pivot(data)
         data2.head()
```

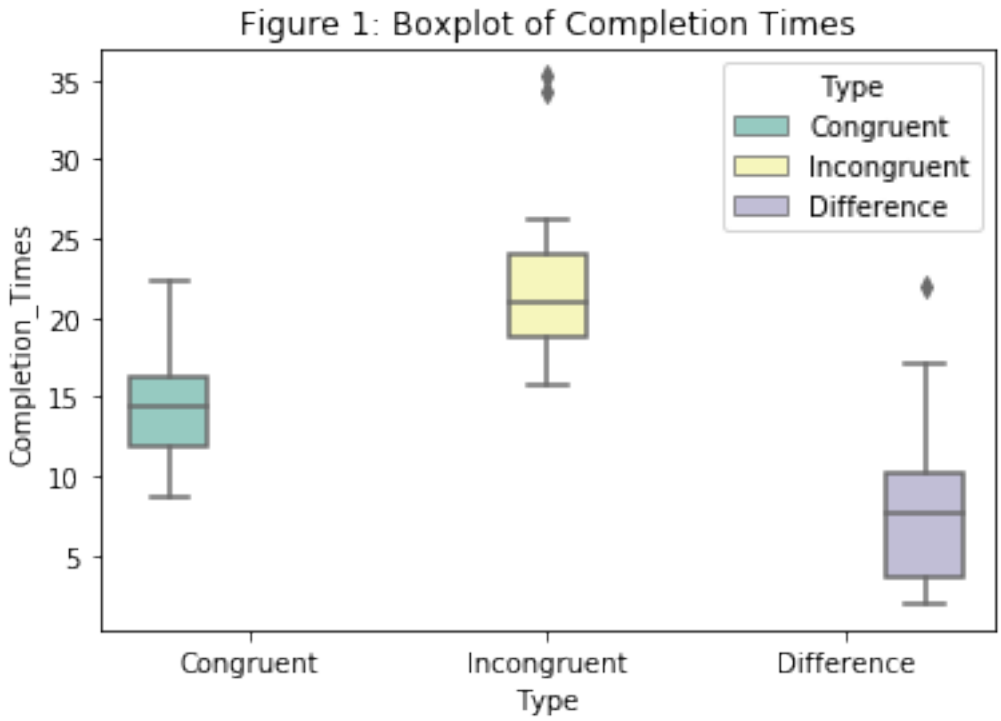
Out[82]:

	Observations	Type	Completion_Times
0	0	Congruent	12.079
1	1	Congruent	16.791
2	2	Congruent	9.564
3	3	Congruent	8.630
4	4	Congruent	14.669

1.8.1 Figure 1: Analysis of Spread

```
In [83]: # Visualization Analysis of Spread
         sns.boxplot(x='Type',
                     y= 'Completion_Times',
                     hue= 'Type',
                     data=data2,
                     palette="Set3").set_title('Figure 1: Boxplot of Completion Times')
```

Out[83]: <matplotlib.text.Text at 0x11e14a950>



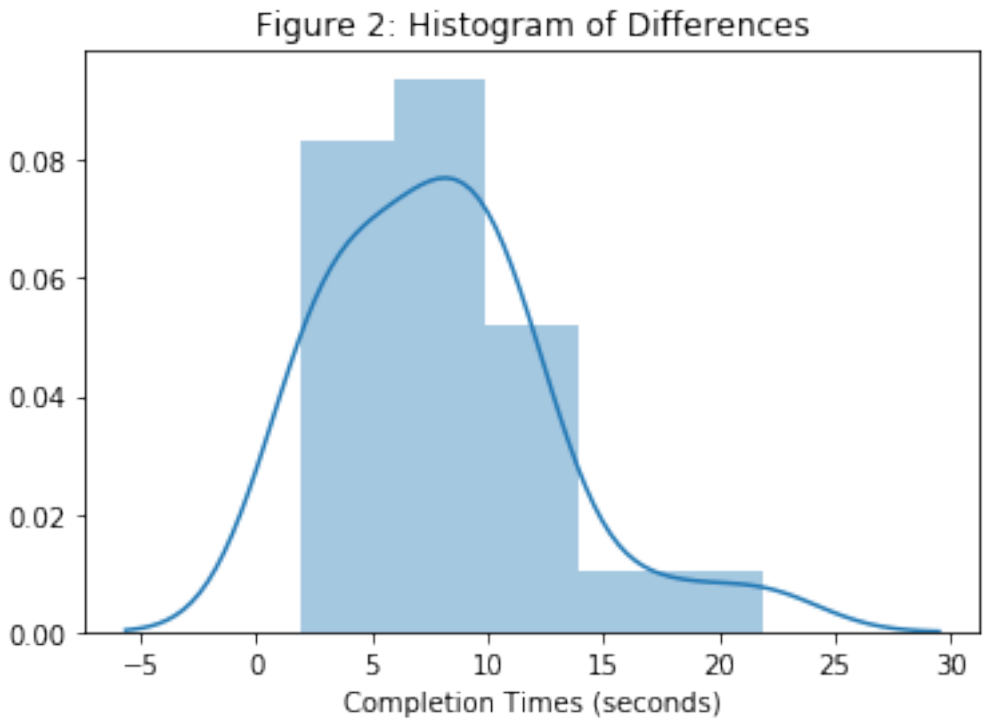
The boxplots for Figure 1 expose the visual representation of the summary statistics from part 3. on the purple box-and-whisker plot, we observe that there is a large upper whisker (top quartile) with a large spread, which translates to a right-skewed probability density function. This is verified in Figure 2, which displays a histogram of the differences set D, with a tendency for a larger spread above the mean. It is also verified by observing Table 2, which verifies that the mean of differences $\bar{d} \approx 7.97$ is larger than its median, around 7.67, which is indicative of a right-skewed distribution.

Figure 1 also exposes the existence of outliers for the incongruent treatment as shown from the box-and-whisker plot of the incongruent set Φ (shown in yellow), which may have influenced the existence of the few outliers in the difference box-and-whisker plot (shown in purple). This will influence our interpretations and the goodness of fit of the data to an approximately normal distribution, which is a significant assumption for performing the parametric paired sample t-test.

1.8.2 Figures 2 and 3: Histograms and Normality Approximations

```
In [15]: sns.distplot(data['Difference'],
                      axlabel='Completion Times (seconds)').set_title('Figure 2: Histogram of Di
```

Out[15]: <matplotlib.text.Text at 0x11b689510>




```
In [16]: for i in ['Congruent','Incongruent', 'Difference']:
             sns.distplot(data[i], label = i, axlabel='Completion Times (seconds)').set_title(
                 'Figure 3: Histogram of Treatment Types')
         plt.legend()
```

```
Out[16]: <matplotlib.legend.Legend at 0x11dcac690>
```

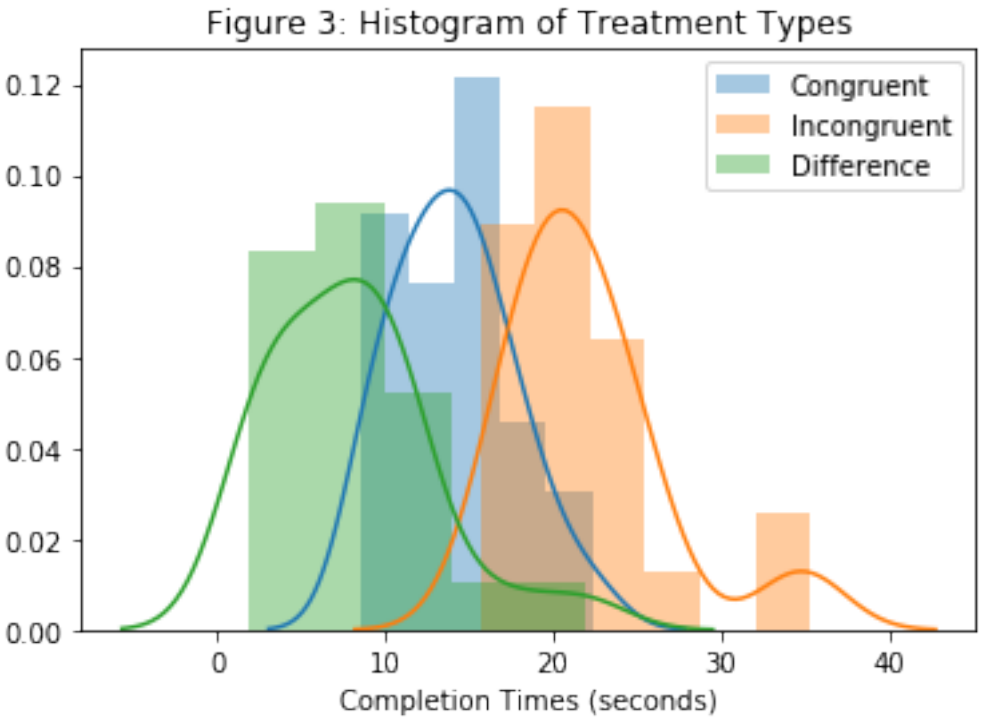


Figure 3 displays all the histograms of congruent and incongruent treatments with their difference, which shows that the difference is the leftmost shifted plot which may indicate that the average differences between the two treatments ("$\Phi - \Theta$" so to speak) are not too drastically different to result in statistically significant results.

1.8.3 Goodness of Fit Test: Kurtosis Test

Figures 1 and 2 display that the data may not approximate a normal distribution very well due to an evident skew to the right, yet since we have a small sample size and at least two outliers, this is unclear. We will perform a more quantitative test called the kurtosis test, a hypothesis test from the SciPy package whose null asserts that the data is of a normal distribution[9].

If we compare with an alpha level $\alpha_{kurtosis} = 0.05$, and observe the test below:

```
In [17]: stats.kurtosistest(data['Difference'])

Out[17]: KurtosistestResult(statistic=1.6441688335149749, pvalue=0.10014133279077567)
```

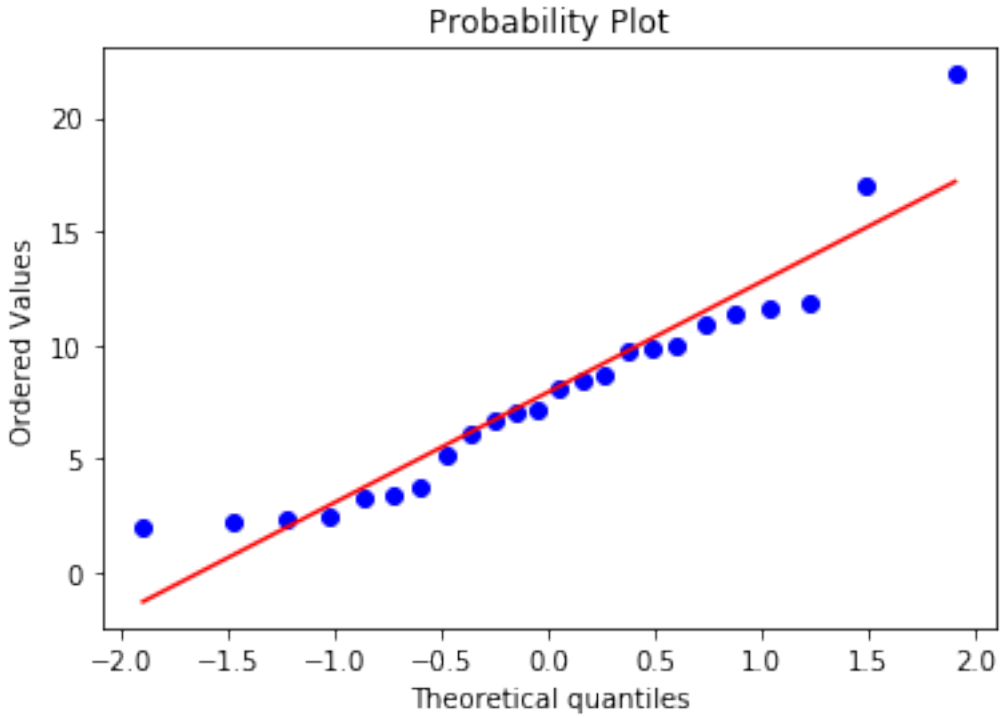
Then with a $p_{kurtosis} \approx 0.1001 > \alpha_{kurtosis}$, we retain and do not reject the null hypothesis which is in itself inconclusive.

1.8.4 Figure 4: Goodness of Fit Test: Probability Plot

```
In [27]: print("Figure 4: Probability Plot")
         stats.probplot(x = data['Difference'], dist='norm', plot = plt)
```

Figure 4: Probability Plot

```
Out[27]: ((array([-1.90380091, -1.48287381, -1.22601535, -1.03156092, -0.8698858 ,
                  -0.7282709 , -0.59996024, -0.48085763, -0.36822879, -0.26009875,
                  -0.154935  , -0.05146182,  0.05146182,  0.154935  ,  0.26009875,
                   0.36822879,  0.48085763,  0.59996024,  0.7282709 ,  0.8698858 ,
                   1.03156092,  1.22601535,  1.48287381,  1.90380091]),
           array([  1.95 ,   2.196,   2.348,   2.437,   3.346,   3.401,   3.727,
                    5.153,   6.081,   6.644,   7.057,   7.199,   8.134,   8.407,
                    8.64 ,   9.79 ,   9.88 ,  10.028,  10.95 ,  11.361,  11.65 ,
                   11.802,  17.055,  21.919])),
          (4.848714513731152, 7.9647916666666658, 0.95231198471853418))
```



We then refer to Figure 4, a probability plot for further inspection of the goodness of fit to a normal distribution. A probability plot is a scatter plot of the theoretical quantiles of the expected distribution plotted against the ordered values of the observed distribution[7].

Here, our observed distribution, plotted in the Y-axis, is the difference set D and the expected distribution, whose quantiles are plotted in the X-axis, is the normal distribution as specified by the `dist = norm` scipy function argument. An observed set of data that closely follows a normal distribution would have a linear relationship in the plot. Sans outliers, we see an approximately linear relationship. The tails of the plot may be a bit different: the data points on the lower-left quadrant indicate that the observed values on the left are larger than that of the expected normal distribution, which would lessen the degree of fitting to a normal distribution. There are also a few

points on the upper-right quadrant (except the outliers) that are beneath the red line, indicating that those observed values are lower in value than a theoretical normal distribution.

1.8.5 Concluding Interpretations from Visualizations

However for the most part, the data follows a linear relationship between observed data and normal distribution quantiles, which favors the argument towards a data that well approximates a normal distribution. We can attribute its skewness due to outliers as well as a small sample. As such, with a fail-to-reject kurtosis test, skewed histogram and box-and-whisker, and approximately linear probability plot, the data can be assumed to be approximately normally distributed for its sample size.

We do have outliers, yet they are not significantly far from the rest of the data, so we will assert that Assumptions 3 and 4 are met in the Statistical Test Formulation section (2b) and we will go ahead and assume that the paired sample t-test is a sufficient test to use for this data.

1.9 5. Statistical Test Implementation

With a statistical test in hand, we can then perform the necessary steps towards finding a test statistic, computing a p-value, and making a decision with stated null and alternate hypotheses. We restate the hypotheses again and other necessary pieces of the paired 2-sample t-test:

1.9.1 Null and Alternate Hypotheses:

$$H_0 \ : \ \mu_d = 0$$

$$H_1 \ : \ \mu_d > 0$$

1.9.2 Basic Statistics

$\bar{d} = 7.964792$
$s_d = 4.864827$
$n = 24$
$df = 23$
$\alpha = 0.0005$

1.9.3 Estimated Standard Error of the Mean

This is defined as follows:

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{4.864827}{\sqrt{24}} = 0.99302865307622179$$

1.9.4 T-Statistic

We compute the t-statistic for a paired 2-sample t-test with the following formulation[7]:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$$

As such:

$$t = \frac{\bar{d} - \mu_d}{s_d} = \frac{7.964792 - 0}{0.99302865307622179} = 8.0207$$

With degrees of freedom $df = n - 1 = 23$[6]. Since this is the case, we find the critical value of a T-distribution with a chosen $\alpha = 0.0005$, we have[10]:

$$T_{\alpha=0.0005} = 3.768$$

1.9.5 P-value

We thus gain a p-value p for the random variable of differences set D distribution, X, with the following definition[6] and implementation:

$$p = \mathbb{P}\{X > t = 8.0207\} = 4.103\text{x}10^{-08}$$

Note that p and t are the same results as that done by the SciPy package function

```
In [44]: stats.ttest_rel(a= data['Incongruent'],
                         b= data['Congruent'])
```

```
Out[44]: Ttest_relResult(statistic=8.020706944109957, pvalue=4.1030005857111781e-08)
```

1.9.6 Conclusion

We can then see that for a right-tailed test, we observe that at an alpha level of $\alpha = 0.0005$, we arrive at a test statistic that is within the critical region of a one-tailed test, which means that we reject the null hypothesis and assert that there is significant evidence that there is a difference in means of completion times between reading color words that match their colors and reading color words that do not match their colors, which shows significant evidence for a treatment effect and for associative inhibition.

1.10 Question 6: Extending the Investigation

While a one-tailed paired-sample t-test proved to be statistically significant, this was based on the assumptions of normality and ignoring two outliers, namely, $d = 17.055, 21.919$. We can extend this investigation by probing if the hypothesis decision will change based off of two different factors:

1. Ignoring outliers
2. Using a non-parametric test

1.10.1 6a. Ignoring Outliers

We will eliminate $d = 17.055, 21.919$ and see if there is a difference in our results.

```
In [87]: new_data = data.sort_values(by= ['Difference'], ascending=True)[::-1]

         new_data.head()
```

```
Out[87]:     Congruent  Incongruent  Difference
         19     12.369       34.288      21.919
         14     18.200       35.255      17.055
         9      14.480       26.282      11.802
         2       9.564       21.214      11.650
         8       9.401       20.762      11.361
```

```
In [88]: new_data = new_data[2:] # Eliminate outliers
         new_data.head()
```

```
Out[88]:     Congruent  Incongruent  Difference
         9      14.480       26.282      11.802
         2       9.564       21.214      11.650
         8       9.401       20.762      11.361
         20     12.944       23.894      10.950
         15     12.130       22.158      10.028
```

Ignoring Outliers: Hypothesis Test Results

```
In [86]: stats.ttest_rel(a= new_data['Incongruent'],
                         b= new_data['Congruent'])
```

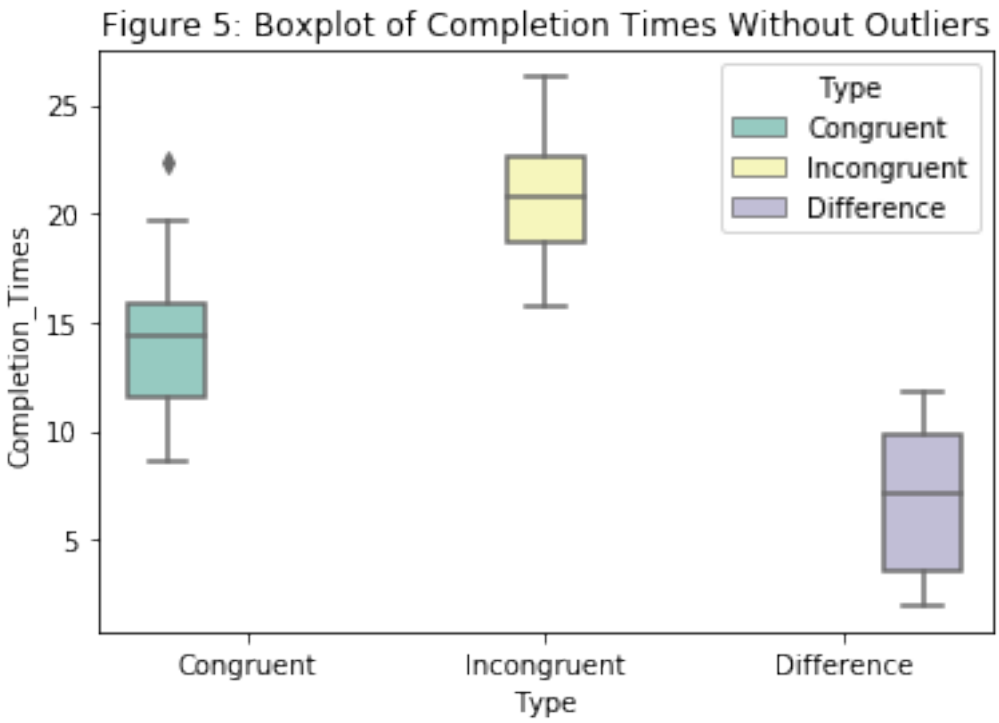
```
Out[86]: Ttest_relResult(statistic=8.6670463737765679, pvalue=1.2063970794825227e-07)
```

The parametric paired 2-sample t-test still yields a test statistic and p-value p that are significant within the critical region determined by $\alpha = 0.0005$. As such, we still conclude to reject the null hypothesis and assert that there is a significant difference between treatments of congruent and incongruent task completion times, yielding for a strong argument for associative inhibition in effect. However, this all relies on the assumption that the clipped data without outliers, which we will denote as $\hat{D}$, follows a normal distribution. We will again test the goodness of fit with visualizations and the kurtosis test:

Ignoring Outliers: Goodness of Fit Visualizations

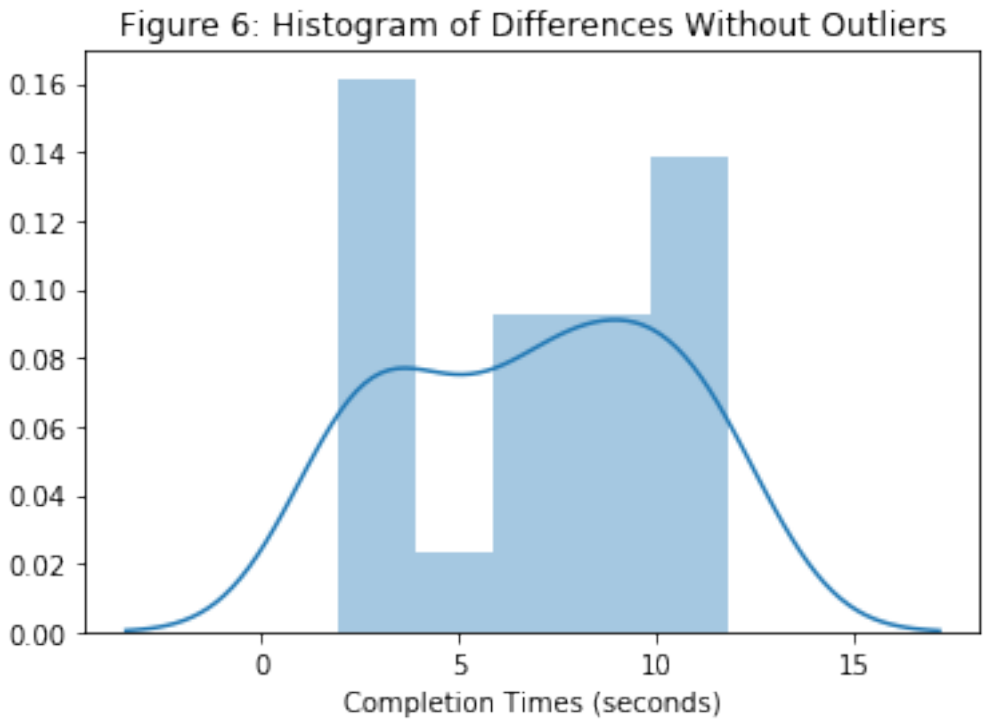
```
In [92]: # Visualization Analysis of Spread without Outliers
         new_data_boxplot = pivot(new_data)
         sns.boxplot(x='Type',
                     y= 'Completion_Times',
                     hue= 'Type',
                     data=new_data_boxplot,
                     palette="Set3").set_title('Figure 5: Boxplot of Completion Times Without Ou
```

```
Out[92]: <matplotlib.text.Text at 0x11f3717d0>
```




```
In [90]: sns.distplot(new_data['Difference'],
                      axlabel='Completion Times (seconds)',
                     bins=5).set_title('Figure 6: Histogram of Differences Without Outliers')

Out[90]: <matplotlib.text.Text at 0x11f19f4d0>
```



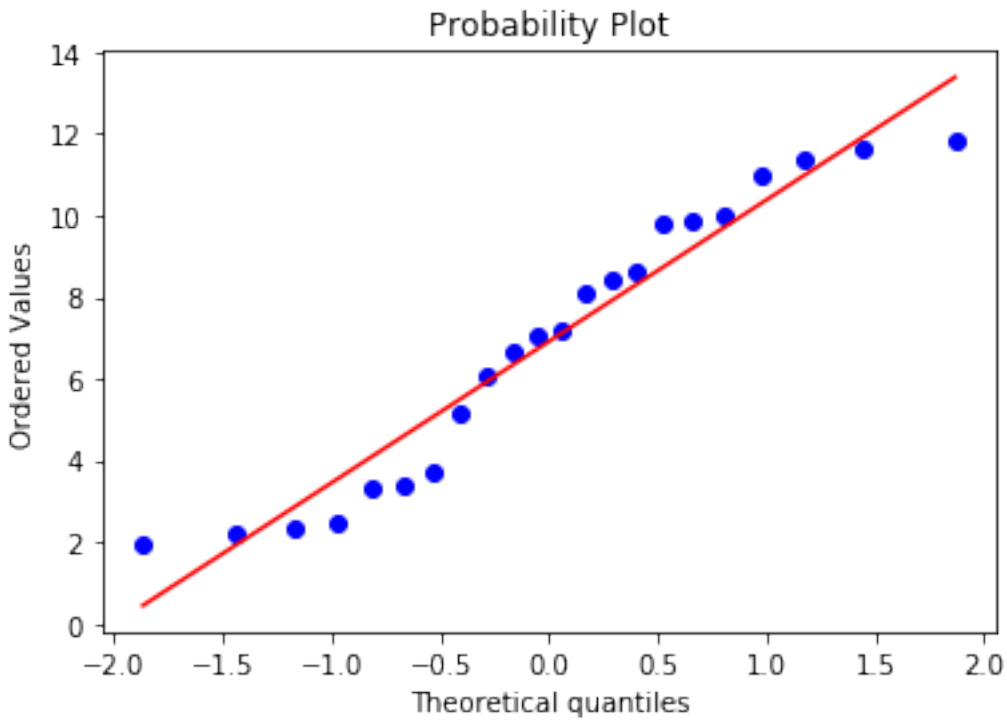

```
In [70]: stats.kurtosistest(new_data['Difference'])
```

```
Out[70]: KurtosistestResult(statistic=-2.3350411265825342, pvalue=0.019541272381413222)
```

```
In [89]: print("Figure 7: Probability Plot Without Outliers")
         stats.probplot(x = new_data['Difference'], dist='norm', plot = plt)
```

```
Figure 7: Probability Plot Without Outliers
```

```
Out[89]: ((array([-1.86607372, -1.43791453, -1.17527742, -0.97550628, -0.80861848,
                  -0.66170536, -0.52787584, -0.40291473, -0.28396605, -0.16891711,
                  -0.05606845,  0.05606845,  0.16891711,  0.28396605,  0.40291473,
                   0.52787584,  0.66170536,  0.80861848,  0.97550628,  1.17527742,
                   1.43791453,  1.86607372]),
           array([  1.95 ,   2.196,   2.348,   2.437,   3.346,   3.401,   3.727,
                    5.153,   6.081,   6.644,   7.057,   7.199,   8.134,   8.407,
                    8.64 ,   9.79 ,   9.88 ,  10.028,  10.95 ,  11.361,  11.65 ,
                   11.802])),
          (3.4683905111296935, 6.9173181818181808, 0.97157809020385688))
```



The box-and-whiskers plot in Figure 5 yields little indicaiton of difference, yet more can be said with Figure 6, 7, and the kurtosis test. Here, for $\alpha = 0.05$, we have a $p_{kurtosis} = 0.01$, which is in the critical region for said α, contrary to the $p = 0.1001$ for D, the original difference set with outliers. As such, we can tell that removing 2 outliers will shift us into the rejection region for $\alpha = 0.05$, and showing that the data will have significant evidence for rejecting the kurtosis null hypothesis that $\hat{D} \sim \mathcal{N}(\mu_d, \sigma_d^2)$ (normally distributed) with population variance of differences σ_d^2, so $\hat{D}$ is not normally distributed from that test. Figure 6 also shows a histogram whose bins seem quasi-bimodal, and the probability plot in Figure 7 verifies a non-linear relationship with the expected distribution being normal. As such, all three of these goodness-of-fit tests fail to support a strong fit that $\hat{D}$ is normally distributed, and as such fail one of the assumptions of a dependent sample t-test. We then revert to a non-parametric test to see the results.

1.10.2 6b. Non-Parametric Test: Wilcoxon Signed Rank Test

Due to the results in 6a, we will inspect the differences between congruent and incongruent treatments D with a non-parametric test. We will first investigate the assumptions to see if we can qualify to implement the statistical test on our data, specifically for a type of non-parametric test called a **Wilcoxon Signed Rank Test**[11]:

Assumptions: *1. For every d_i, ϕ_i and θ_i are dependent on one another.*

This is true, since both ϕ_i and θ_i, the incongruent and congruent completion times for each observation, is dependent on the observation that took those times.

2. Every participant is independent of one another.

This is true, because we assume that no two participants influenced each others scores and we randomly selected participants for this study.

3. The set D has to be a continuous dependent variable.

Since D is a completion time measurement and time is on a continuous space, so is D.

4. The measurements from D are ordinal (can be ordered).

This is true. For $D \subset \mathbb{R}$, every element can be ordered, as we did with the `new_data` variable.

Since the assumptions are all met, we can proceed with the statistical test. We explain the algorithm and intuition behind the test as follows[7]:

Test Procedure

1. Calculate the differences, d_i , and the absolute values, $|d_i|$, of the differences and rank the latter into a variable R.
2. Exclude pairs where $d_i = 0$. As such, the cardinality of D would be reduced to n_r, where $n_r < n$.
3. Restore the signs of the differences to the ranks, denoted R, obtaining signed ranks, denoted R_s.
4. Calculate W_+, the sum of those ranks that have positive signs, as follows:

$$W = \sum_{i=1}^{n_r} \left([d_i]_+ \cdot R_i \right)$$

where $[d_i]_+ \doteq d_i \in D : d_i > 0$.

4. Calculate W_- this way, but with the negative signed ranked differences. Choose the smaller of the two sums, $W^* = min(W_+, W_-)$ and compare with the critical value for a pre-computed distribution of Wilcoxon critical values, based on number of tails in test, sample size, and α value.

Intuition From John Rice's *Mathematical Statistics and Data Analysis* text[7]:

> The idea behind the signed rank test (sometimes called the Wilcoxon signed rank test) is intuitively simple. If there is no difference between the two paired conditions, we expect about half the d_i to be positive and half negative, and W_+ will not be too small or too large. If one condition tends to produce larger values than the other, W_+ will tend to be more extreme. We therefore can use W_+ as a test statistic and reject for extreme values.

The signed rank test allows us to focus on the ordinal relationships of data points, which allows it to be resistant to outliers, one of the violations in our assumptions for a parametric paired t-test. Since it does not assume normality, we will have to draw it from a tested null distribution to test if the test statistic falls within the critical region to make a decision. For now, we begin with the test implementation, with a chosen value of alpha $\alpha = 0.05$.

Hypothesis Test We then construct the following claim:

H_0 **: median difference is 0, i.e., for the null distribution** F **such that it is the distribution of** d_i**,** F **is symmetric about 0.**

H_1 **: median difference is greater than 0, i.e.,** F **is not symmetric about 0.**

This null distribution, F, is derived from the following idea: + Assume H_0 is true.

- For every observation i, $d_i \sim F$ indicates that d_i are distributed by the null distributed F.
- Since $d_i = \phi_i - \theta_i$, the difference of the incongruent case from the congruent case, d_i and $-d_i$ both come from F due to the assumption of symmetry in the distribution. As such, they have equal probability. We then have 2^{n_r} assignments of rank to calculate W_+, since each W_+ has a probability of $\frac{1}{2^{n_r}}$[7].
- One can calculate the probability null distribution this way, and is available in many computer packages such as SciPy.

```
In [94]: stats.wilcoxon(x= data['Congruent'], y= data['Incongruent'])
```

```
Out[94]: WilcoxonResult(statistic=0.0, pvalue=1.821529714896801e-05)
```

Conclusion Using the SciPy package, we see that our statistic $W^* = W_-$, since there are no negative differences in the data, and so $W^* = W_- = 0$. This is statistically significant against our critical value for $\alpha = 0.05$[12], and as such we reject the null hypothesis and the data is not symmetric about 0.

1.11 Summary

As such, with a vague assumption of normality, we attempted the parametric, paired, dependent sample t-test, achieving a high test statistic and p-value to reject the null hypothesis that $H_0 : \mu_d = 0$, i.e., there is a significant difference between the incongruent and congruent task completion times. However, after removing outliers in our data set of differences D, the new data set $\hat{D}$ did not meet our qualitative and quantitative goodness-of-fit tests, and so we resorted to a non-parametric test, specifically the Wilcoxon signed rank test, which does not assume a normal distribution and is resistant to outliers. After formulating a null hypothesis that the null distribution is symmetric about 0, our test shows that the data and its p-value, $p = 1.8216\text{x}10^{-05}$, is in the critical region of $\alpha = 0.05$, showing significant evidence to reject the null and show a treatment effect for a null distribution not symmetric around 0, which implies that the difference set is most likely positive and that there is a significant difference between incongruent and congruent completion times. As such, both parametric and non-parametric approaches support this decision, which asserts the existence of associative inhibition in trying to assign reading color names with different colors.

1.11.1 References:

1. Stroop Effect Personal Test
2. Stroop Effect Website
3. The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm
4. Stroop and Picture-Word Interference are two sides of the same coin
5. Stroop Original Paper
6. Paired Sample T-Test
7. Rice, John A. *Mathematical Statistics and Data Analysis*. Belmont, CA: Thomson/Brooks/Cole, 2007, pp. 370-359, 444-459.

8. Within-Subject Design
9. Scipy Kurtosis Test Documentation
10. T-Table for One-Tailed and Two-Tailed Tests
11. Wilcoxon Signed Rank Test Assumptions
12. Wilcoxon Signed Rank Test Table

```
In [ ]:
```