Table of Contents

Section 1. Statistical research methods

1.1 Learning outcomes

The following topics were studied and practiced using some practice exercises and applied to the current project

- Construct and Operational Definition
- Population vs. Sample
- parameter vs. statistic
- Experimentation
- Observational study
- Treatment
- Independent and dependent variables
- Treatment group vs. Control group and Placebo
- Blinding and double blinding
- Normal distribution and central tendencies
- Variability – Box plots, Histograms, IQR
- Outliers, Variance, Standard Deviation
    - In a normal distribution 65% of the data lies within 1 standard deviation from the mean, 95% within 2 standard deviations, and 99.7% within 3 standard deviations.
- Bessel's correction
- Probability distribution
- Standardizing - Finding Z-score and learn to use the Z-table
- Central Limit Theorem
- Confidence intervals
    - Margin of error, Critical Z-score
- Hypothesis test
    - Alpha levels, Type I error, Type II error
- T-tests
    - T-distribution, t-statistic vs. z-statistic
    - One tailed & two-tailed t-tests
    - P-value and statistical significance

  - correlation coefficient and the coefficient of determination
  - One sample and two-sample t-tests
  - Dependent samples (Within-Subject Designs -> 2-conditional, longitudinal, pre-test<->post-test)
  - Independent samples (Between Subject Designs)
- Cohen's d to measure the effect size of the strength of a phenomenon
- Standard Error, Pooled Variance
- Conditional probability and Bayesian decision theory

1.2 Project 1: Conduct statistical tests on data and draw conclusions

The objective of this project is to apply the above learnt statistical research methods on three data sets, each time performing a different test and drawing conclusions based on the test. Hence Project 1 is divided into three parts as follows:

1.2.1 Part 1: Determine the optimal length of chopsticks

Objective:

In the past, researchers conducted two laboratory studies to evaluate the effects of the length of the chopsticks on the food-serving performance of adults and children.
The results showed that the length of the chopsticks significantly affected the food-pinching performance, and that chopsticks of about 240 and 180 mm long were optimal for adults and pupils, respectively.

Our goal is to corroborate the above results of the published paper, by using a new experimental sample with 31 male junior college students serving as subjects. The subjects were asked to use chopsticks of lengths of 180, 210, 240, 270, 300, and 330 mm to pickup peanuts from a cup and drop in a second cup. (this calculated as a percentage is called the food pinching performance)

The purpose of this project is to fall at ease into identifying various important terms and operational definitions using a simple dataset and analyze the data to form logical conclusions to corroborate the pre-established paper on chopstick effectiveness.

Data used:

"Chopstick-effectiveness.csv"

Tools used:

Anaconda iPython

Language:

python

Analysis:

STEP 1: IDENTIFICATION OF DEPENDENT AND INDEPENDENT VARIABLES IN THE EXPERIMENT

The independent variable in this experiment is the "Chopstick.Length"
The dependent variable (or the outcome) of the experiment is the "Food.Pinching.Efficiency" of the chopsticks

STEP 2: FORMULATE THE OPERATIONAL DEFINITION

The dependent variable, which is the "Food.Pinching.Efficiency" of the chopsticks is operationally defined as follows: The percentage of peanuts picked and placed in a cup (PPPC) by a pair of chopsticks. (Percentage of The number of peanuts successfully placed out of the total number of peanuts provided)

Step 3: Identify the Controlled variables

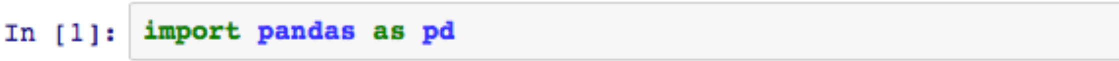
Based on the description of the experiment and the data set, we can identify the following as at least 3 different factors that were controlled in the experiment:

a) *The Gender of the participants* - All the participants are male

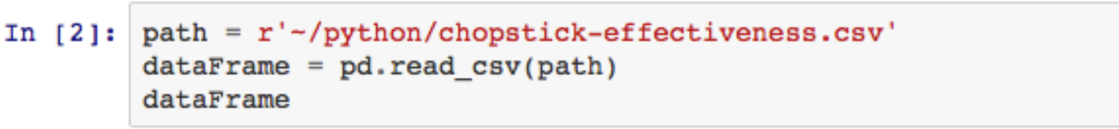
b) *The material and Grip of the chopsticks* - All the chopsticks given to participants is of the same type to maintain unbiased testability

c) *The age group of the participants* - Since all the adult participants of part 1 of the experiment belonged to junior colleges, they are more likely to be in a fixed age group.

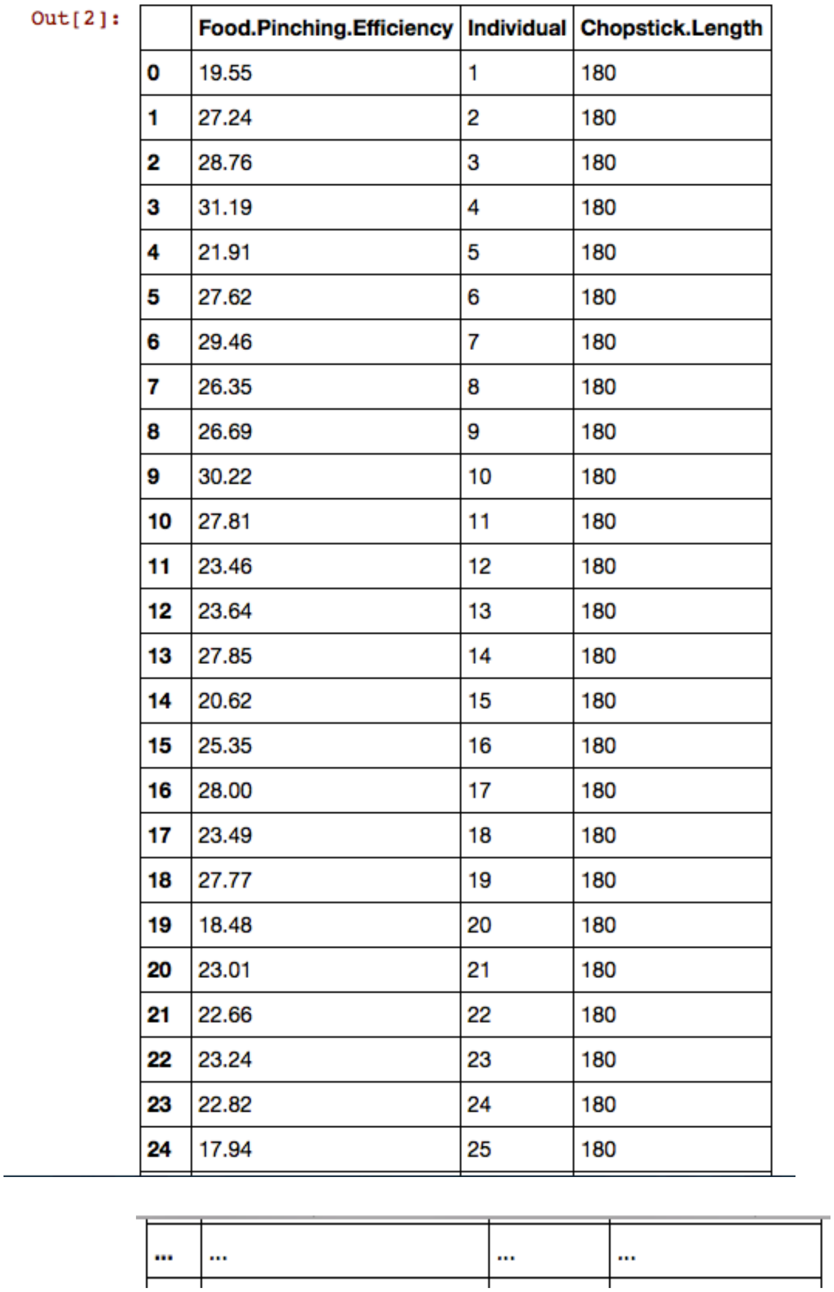
Now lets look at the data using python pandas library



```
In [1]: import pandas as pd
```

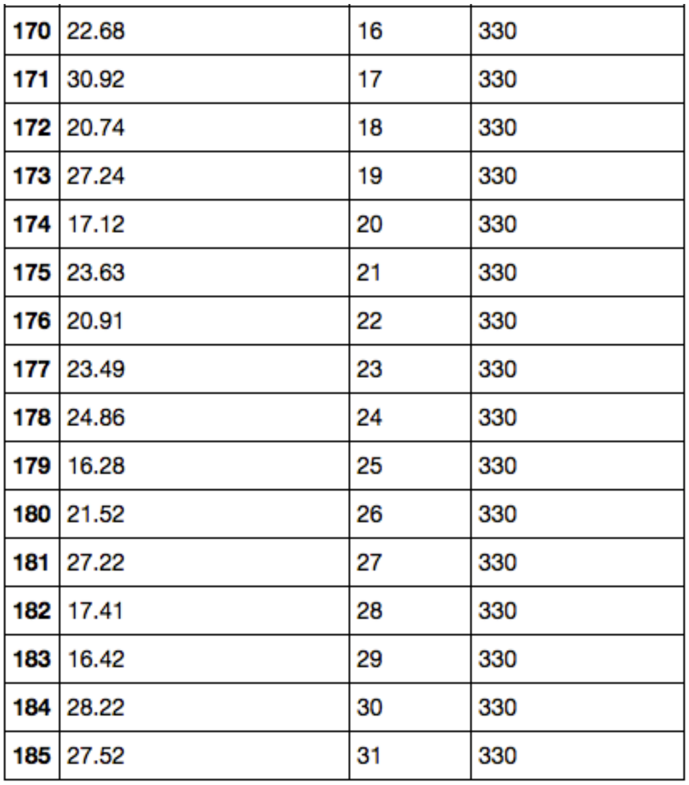


```
In [2]: path = r'~/python/chopstick-effectiveness.csv'
        dataFrame = pd.read_csv(path)
        dataFrame
```



Out[2]:

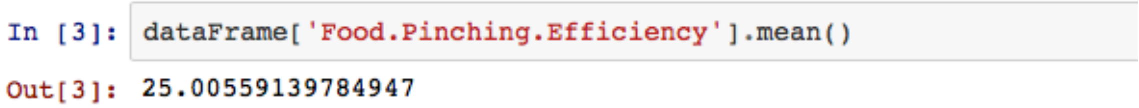
	Food.Pinching.Efficiency	Individual	Chopstick.Length
0	19.55	1	180
1	27.24	2	180
2	28.76	3	180
3	31.19	4	180
4	21.91	5	180
5	27.62	6	180
6	29.46	7	180
7	26.35	8	180
8	26.69	9	180
9	30.22	10	180
10	27.81	11	180
11	23.46	12	180
12	23.64	13	180
13	27.85	14	180
14	20.62	15	180
15	25.35	16	180
16	28.00	17	180
17	23.49	18	180
18	27.77	19	180
19	18.48	20	180
20	23.01	21	180
21	22.66	22	180
22	23.24	23	180
23	22.82	24	180
24	17.94	25	180
...	...	...	...



170	22.68	16	330
171	30.92	17	330
172	20.74	18	330
173	27.24	19	330
174	17.12	20	330
175	23.63	21	330
176	20.91	22	330
177	23.49	23	330
178	24.86	24	330
179	16.28	25	330
180	21.52	26	330
181	27.22	27	330
182	17.41	28	330
183	16.42	29	330
184	28.22	30	330
185	27.52	31	330

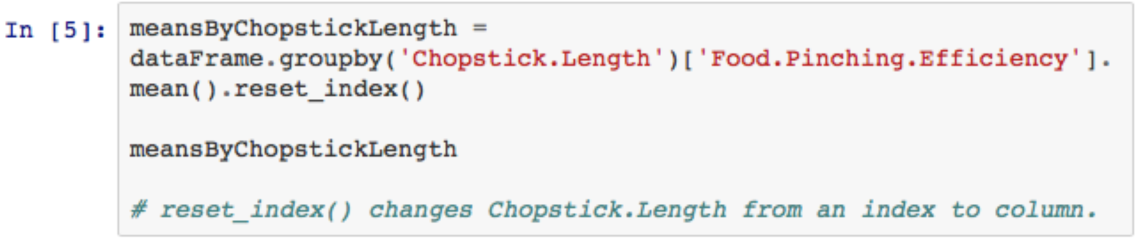
186 rows × 3 columns

Let's calculate the average "Food Pinching Efficiency" for all 31 participants and all chopstick lengths.

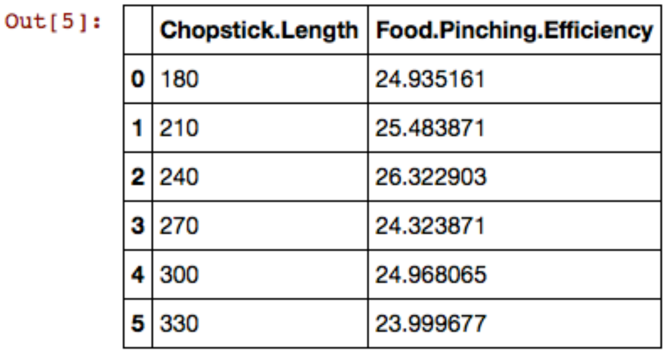


```
In [3]: dataFrame['Food.Pinching.Efficiency'].mean()

Out[3]: 25.00559139784947
```

As we can see, this number doesn't let us know which of the chopstick lengths performed best for the thirty-one male junior college students. Let's break down the data by chopstick length. Let's generate the average "Food Pinching Effeciency" for each chopstick length



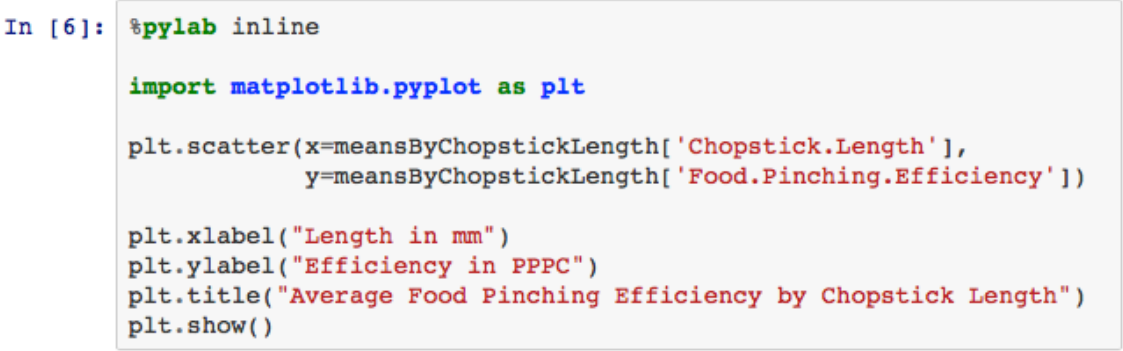
```
In [5]: meansByChopstickLength =
        dataFrame.groupby('Chopstick.Length')['Food.Pinching.Efficiency'].
        mean().reset_index()

        meansByChopstickLength

        # reset_index() changes Chopstick.Length from an index to column.
```



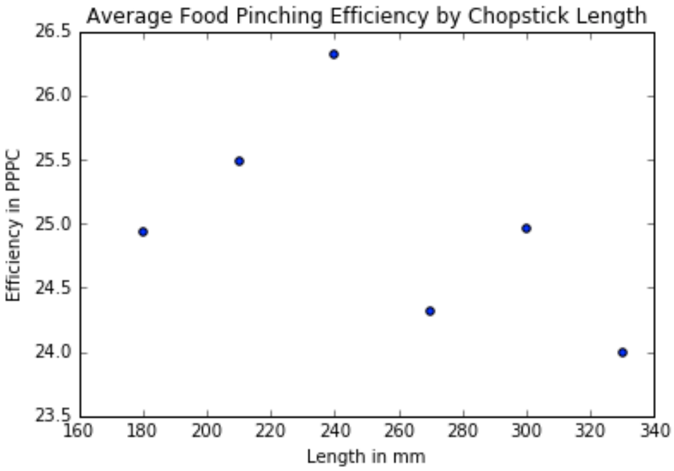
Out[5]:

	Chopstick.Length	Food.Pinching.Efficiency
0	180	24.935161
1	210	25.483871
2	240	26.322903
3	270	24.323871
4	300	24.968065
5	330	23.999677

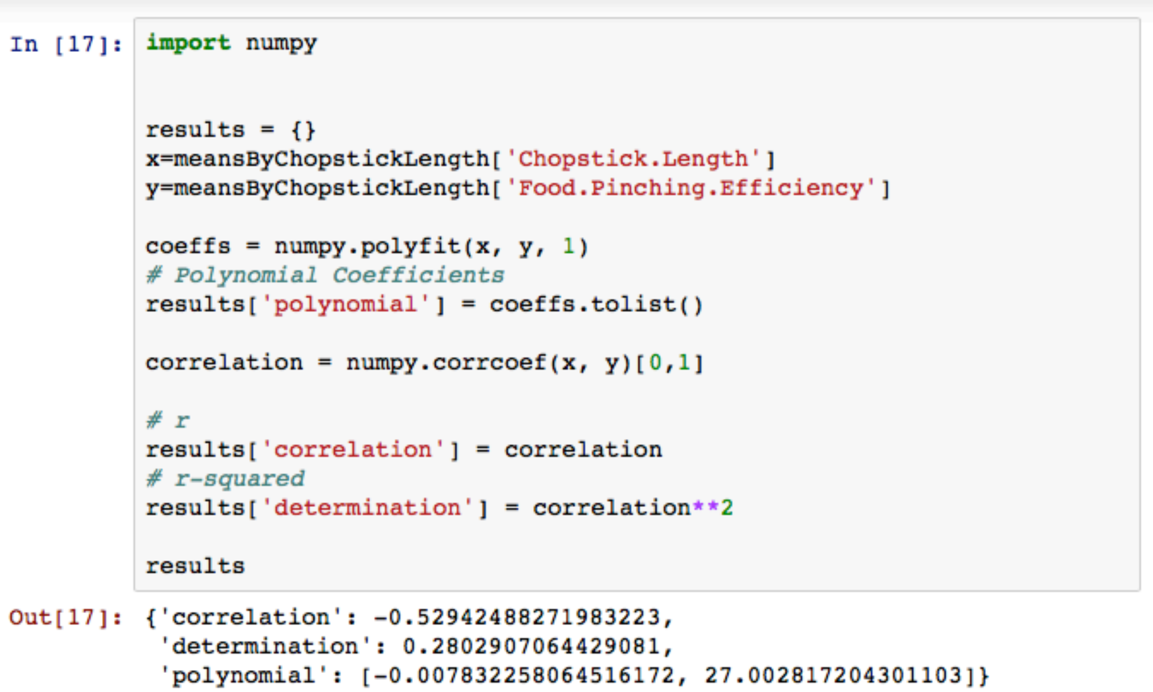
Here, we see the length wise mean performance. Lets plot the results.



```
In [6]: %pylab inline

        import matplotlib.pyplot as plt

        plt.scatter(x=meansByChopstickLength['Chopstick.Length'],
                    y=meansByChopstickLength['Food.Pinching.Efficiency'])

        plt.xlabel("Length in mm")
        plt.ylabel("Efficiency in PPPC")
        plt.title("Average Food Pinching Efficiency by Chopstick Length")
        plt.show()
```



Now, lets, fit a polynomial regression model to this data and see if we get a strong correlation:



```
In [17]: import numpy


         results = {}
         x=meansByChopstickLength['Chopstick.Length']
         y=meansByChopstickLength['Food.Pinching.Efficiency']

         coeffs = numpy.polyfit(x, y, 1)
         # Polynomial Coefficients
         results['polynomial'] = coeffs.tolist()

         correlation = numpy.corrcoef(x, y)[0,1]

         # r
         results['correlation'] = correlation
         # r-squared
         results['determination'] = correlation**2

         results

Out[17]: {'correlation': -0.52942488271983223,
          'determination': 0.2802907064429081,
          'polynomial': [-0.007832258064516172, 27.002817204301103]}
```

We see that the coefficient of determination i.e., r^2 value is 0.28 which only means that 28% of variability in the food pinching efficiency is explained by the length of the chopstick under the chosen model (which is a 2nd degree polynomial regression)

STEP 4: INTERPRETATION OF RESULTS:

Based on the generated scatterplot and the correlation and coefficient of determination values, we can interpret that the "Food pinching efficiency" seems to increase as the chopstick.length increases up to a certain extent and then, beyond the optimum length, the efficiency seems to follow a down hill. There is clearly a parabolic relation between the Efficiency and the Length. This relation shows a single most optimal efficiency giving the corresponding optimal length value.

STEP 5: DRAW CONCLUSIONS:

In the abstract the researchers stated that their results showed the length of the chopsticks significantly affected food-pinching performance, and that chopsticks of about 240 mm long were optimal for adults.

Based on our new data that we analyzed, it seems reasonable to conclude that 240mm is the optimal length of chopsticks for adults.

1.2.2 Part 2: Test a perceptual phenomenon

Background:

For this project, our aim is to investigate a classic phenomenon from experimental psychology called the "Stroop Effect.

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: for example RED, BLUE. In the *incongruent words* condition, the words displayed are color words whose names do not match the colors in which they are printed: for example PURPLE, ORANGE. 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.

Objective:

The objective of this project is to conduct an own observational study on a set of participants who take the stroop test and create a hypothesis regarding the outcome of the task and eventually reject or fail to reject the null hypothesis using the statistical methods learnt earlier.

Data used:	***Tools used:***
"stroopdata.csv"	MS Excel

Analysis:

STEP 1: IDENTIFICATION OF DEPENDENT AND INDEPENDENT VARIABLES IN THE EXPERIMENT

Dependent variable:

The time taken by the participants to read the words for each category of words

Independent variable:

The type of the words in the stroop effect, read by the participants (the congruent words and the incongruent words). The type of the word is the independent variable here

STEP 2: FORMULATE THE HYPOTHESIS (NULL AND ALTERNATE)

Null Hypothesis:

Two-tail

The null hypothesis is that the congruent and incongruent samples come from the same general population. i.e., even though we see a clear difference in sample means in the collected data, it could be by chance and there is actually no difference in response times between the two conditions of the experiment for the population.

$$H_0 : \mu_{congruent} = \mu_{incongruent} \text{ (or) } \mu_{congruent} - \mu_{incongruent} = 0$$

One-Tail

We can also go further and define a one tailed test hypothesis that the Incongruent population has a lower mean time than the Congruent population.

$H_0 : \mu_{congruent} \geq \mu_{incongruent}$ (or) $\mu_{congruent} - \mu_{incongruent} \geq 0$

Alternate Hypothesis:

Two-tail

The alternative hypothesis is that there is a difference between the population-means of the congruent and incongruent, however, we are not assuming which is larger or smaller. The difference in sample means we are witnessing is representative of the general population in that they are different.

$H_A : \mu_{congruent} \neq \mu_{incongruent}$ (or) $\mu_{congruent} - \mu_{incongruent} \neq 0$

One-Tail

The alternative hypothesis is that the population-mean of the incongruent times is greater than the population mean of the congruent times, also meaning that they come from different populations and that the difference in sample means we are witnessing is representative of the general population.

$H_A : \mu_{congruent} < \mu_{incongruent}$ (or) $\mu_{congruent} - \mu_{incongruent} < 0$

STEP 3: CHOOSE A STATISTICAL TEST

There are three major factors in this data that drives us in our decision of choosing the apt statistical test:

(1) We only have samples, we do not have any population parameters (we need the population standard deviation to compute the z-statistic usually)

(2) Our participants from one sample are used for the second sample – suggests a dependents sample test

(3) our data samples are really small

Owing to the above factors, we will choose a paired-samples Dependent T-test for this project

We will choose a "*two-tailed paired-sample dependent T-sample test*" to reject or accept the null-hypothesis.

STEP 4: CALCULATE THE SAMPLE STATISTICS AND THE T-STATISTIC

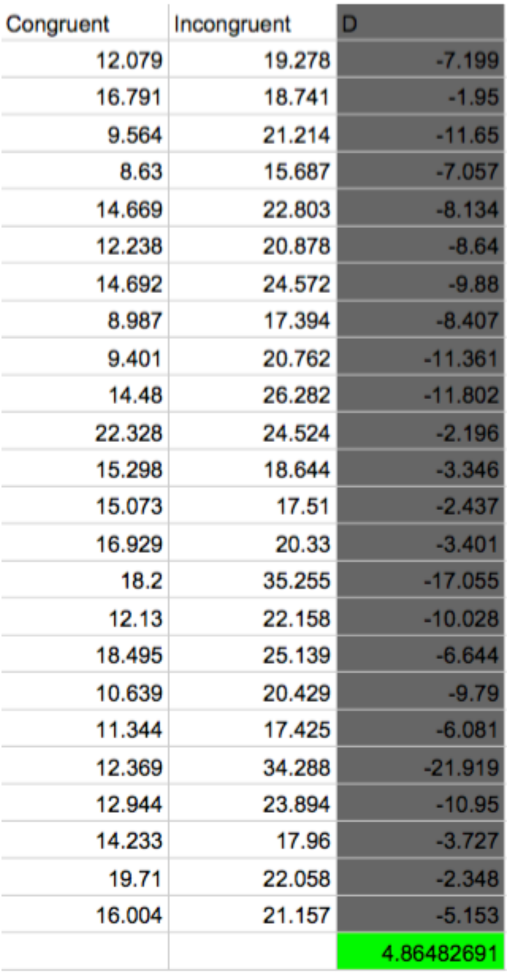
From the data, we can compute the mean of the congruent and incongruent samples

Sample mean of Congruent X_C= 14.051125

Sample mean of incongruent mean X_I= 22.01591667

Sample standard deviation of the difference(D) S_D = 4.86482691

This calculated using the stdev() function in excel that gives us the sample estimate as using SQRT(($\sum$(x_i-x)^2)/n-1)



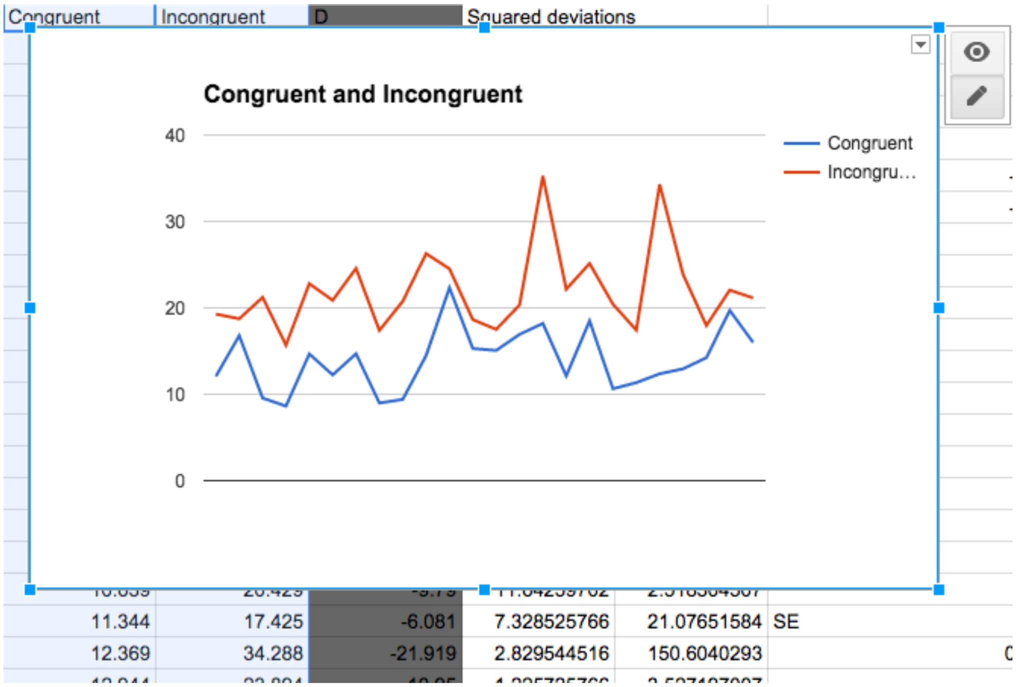
Congruent	Incongruent	D
12.079	19.278	-7.199
16.791	18.741	-1.95
9.564	21.214	-11.65
8.63	15.687	-7.057
14.669	22.803	-8.134
12.238	20.878	-8.64
14.692	24.572	-9.88
8.987	17.394	-8.407
9.401	20.762	-11.361
14.48	26.282	-11.802
22.328	24.524	-2.196
15.298	18.644	-3.346
15.073	17.51	-2.437
16.929	20.33	-3.401
18.2	35.255	-17.055
12.13	22.158	-10.028
18.495	25.139	-6.644
10.639	20.429	-9.79
11.344	17.425	-6.081
12.369	34.288	-21.919
12.944	23.894	-10.95
14.233	17.96	-3.727
19.71	22.058	-2.348
16.004	21.157	-5.153
		4.86482691

Standard Error of mean of difference = S_D/ √n , where n is the size of the sample
Therefore SE_D = 4.86482691/√24 = 0.9930286348

t-Statistic = (X_C - X_I) / SE = -7.964791667/0.9930286348 = **-8.020706944**

STEP 5: SOME VISUALIZATION

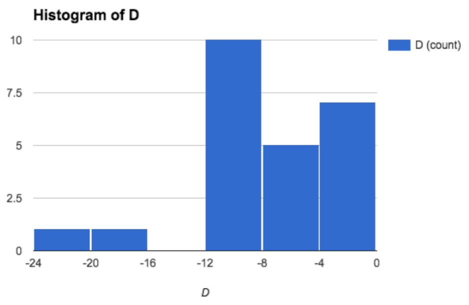
The plot between the sample of Congruent data and incongruent data is shown below:

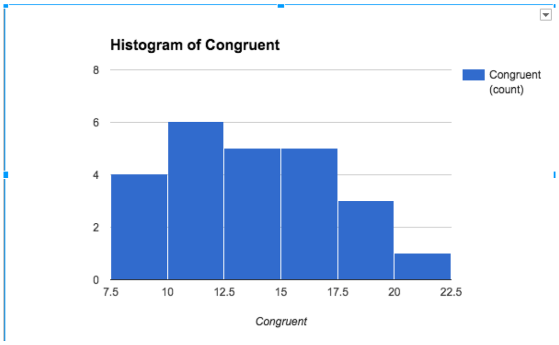


Congruent	Incongruent	D	Squared deviations		
11.344	17.425	-6.081	7.328525766	21.07651584	SE
12.369	34.288	-21.919	2.829544516	150.6040293	

From the above plot, it can be noted that the congruent condition times for every observation falls below the incongruent condition times, for the samples.

Also if we look at the histograms of Congruent and incongruent separately, we can see that they are both normal distributions. And when we subtract the one normal distribution from other, we end up getting a normal distribution





So the column “D” on which we will perform our t-test is also essentially normally distributed.

Step 6: Define confidence interval and calculate the t critical value

The t-statistic as calculated earlier is: -8.021

We can choose to analyze the result at 95% confidence interval in which case our alpha (œ) <0.05(or 5%). We can fetch our t critical values from the t-table.

Since this is a two-tailed test, our tail probability in each tail will be 0.025
Our degrees of freedom (df) = sample size – 1 = 24 -1 = 23

Hence our important factors are as follows:
t-statistic = -8.021
Confidence interval = 95%
Critical œ = 0.05 (or 5%)
Tail probability = 0.025
Degrees of freedom = 23

Now we can fetch our t-critical values at the tails

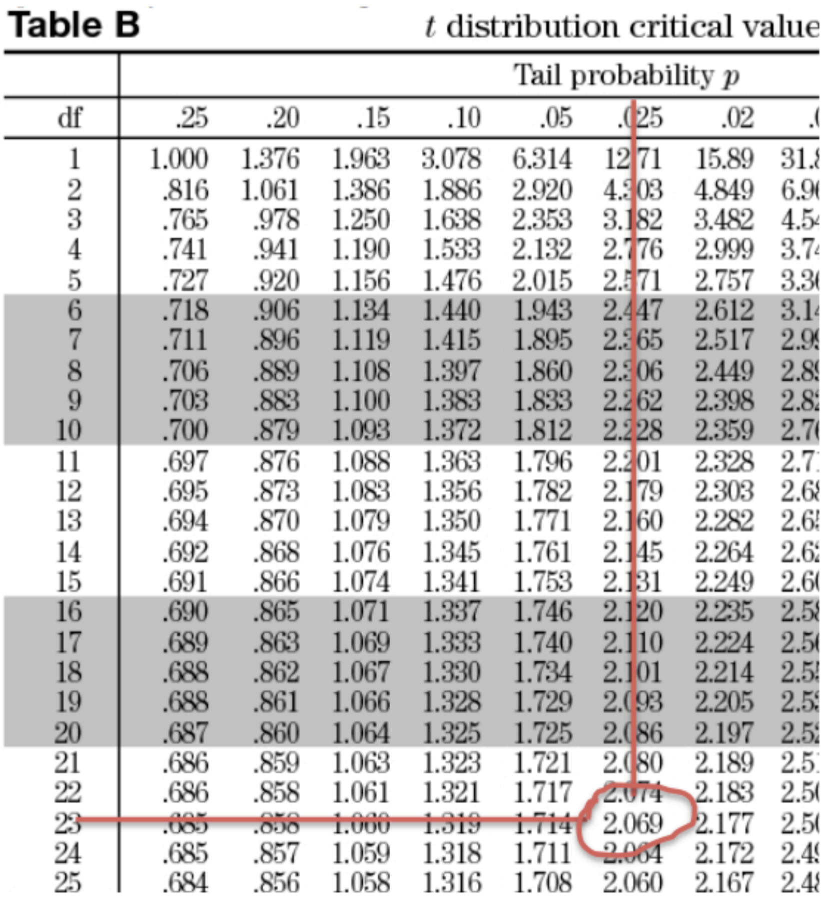


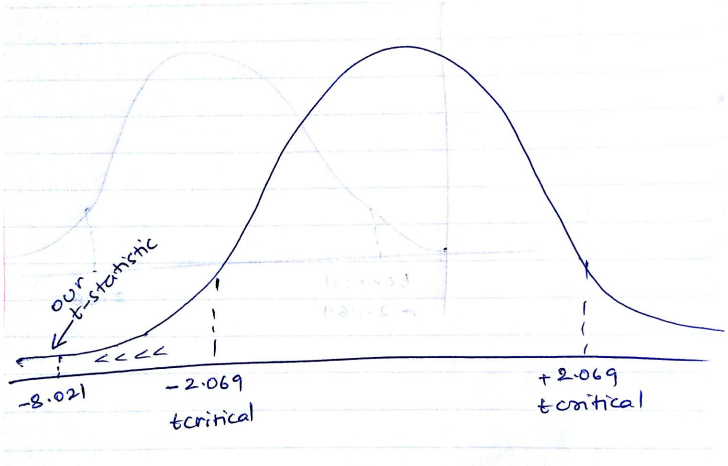
Table B *t* distribution critical value

	Tail probability *p*							
df	.25	.20	.15	.10	.05	.025	.02	.0
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.8
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.9
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.5
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.7
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.3
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.1
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.9
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.8
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.8
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.7
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.7
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.6
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.6
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.6
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.6
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.5
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.5
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.5
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.5
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.5
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.5
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.5
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.5
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.4
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.4

Therefore t Critical = plus or minus 2.069

STEP 7: INTERPRET THE RESULTS AND DRAW CONCLUSION ON THE HYPOTHESIS

If |t statistic| value > t critical value, then we know that our sample's t statistic lies in the critical. (we are ignoring signs here since we have chosen a two-tailed test)
|t statistic| = 8.021; t critical = 2.069
As shown in the below drawn depiction, our t-statistic lies far down in the critical region.



This means that the probability of finding such a difference in the two sample means is <0.05 and hence it is not possible that it occurred by chance.

This gives us grounds to reject the null hypothesis that states that the condition of the word has no effect on the time to read the word

We can conclude that the condition of the word does have an impact on the time taken to read the word (i.e $\mu_C - \mu_I \neq 0$)

To be even clearer on the direction of inequality, from the above depiction, we can see that the t-statistic is a –ve value and lies in the lower tail of the critical region. So, we can say that the difference between population means of the two conditions is less than 0 and so $\boldsymbol{\mu_C < < \mu_I}$***.***

This proves that the fact that almost every participant in the experiment recorded a greater time for incongruent words than the congruent words was not just coincidental but because there is some kind of intrusion involved. This confirms the original paper, which states what the stroop effect actually describes: "The two areas in our brains that resolve two different aspects of a problem come into conflict with each other while reading the Incongruent words and this conflict results in more time taken by the brain to resolve the problem"