Exploring the Stroop Effect using Python

Yigal Weinstein

November 5, 2015

Introduction

So the stage is set with the following description for the Stroop experiment:

> 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 fisplayed 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. [16]

This project attempts to follow the outline provided in [16] and the project rubrick set forth in [17].

Question 1: Identify variables in the experiment

What is our independent variable? What is our dependent variable? incongruency between the color of a word, and the semantic meaning of the word. As example the word 'RED' is a congruent example where the word 'red' has the color of its semantic meaning, i.e. the color it represents, and 'BLUE' is an incongruent case where the word 'blue' which has the semantic meaning of the color blue is instead the color green. Frome here on out the terms congruent color selection, CCS, and incongruent color selection, ICS, will be used to denote the two conditions the dependent variable may take that of a word representing a color being colored with or with a different color respectively.

The dependent variable is the time, in seconds, it takes to read through an equal sized lists of CCS or ICS words.

Question 2: Establish a hypothesis and statistical test

What is an appropriate set of hypotheses for this task? What kind of statistical test do you expect to perform? Justify your choices.

The null hypothesis, H_0 is the statement that the ICS task will take less or an equal amount of time to the CCS task. The alternative hypothesis, H_A is the statement that the mean population time to read an ICS list will be greater than an equally sized CCS list. The null and alternative hypothesis choice stems from a need to test quantatively the validity and extent of an already well respected phenomenon the Stroop effect. The Stroop effect being defined to mean that cognitive incongruences will lead to an increase in the reaction time of a task.

It is true that this single tailed objective may not provide insight if indeed the reaction time for ICS word lists take, on the average, less time to complete, however this is a test regarding the scope and veracity of the Stroop effect which has been shown in many previous quantative tests to have general validity. If indeed the ICS task takes less time it is enough that H_0 should be kept to cause doubt regarding the validity of either the test or the Stroop effect. That is due to the large body of work already done to test the effect if H_0 is kept there is either something wrong with the experiment or the Stroop effect has either more limitations on generalizability than expected or there are indeed real issues with it's veracity. So to sum things up via a set of equations:

$$H_0: \mu_{ICS} \leq \mu_{CCS} \tag{1}$$
$$H_A: \mu_{ICS} > \mu_{CCS} \tag{2}$$

As the two sets of data are correlated, the same subjects are used to take both the ICS and CCS tests, the proper test to use is a dependent t-test for paired samples. Given the hypothesis devised above the single tail test will be used.

Question 3: Report descriptive statistics

Report some descriptive statistics regarding this dataset. Include at least one measure of central tendency and at least one measure of variability.

In Figure 1 is printed the mean, sample standard deviation, minimum values, 25th, 50th - median, 75th percentiles, and maximums for both the CCS and ICS samples, all values are in seconds.

	Congruent	Incongruent
count	24.000000	24.000000
mean	14.051125	22.015917
std	3.559358	4.797057
min	8.630000	15.687000
25%	11.895250	18.716750
50%	14.356500	21.017500
75%	16.200750	24.051500
max	22.328000	35.255000

Figure 1: The source code for the table may be found: p1_stroop_effect.py. The above table was generated with help from official PANDAS API documentation as follows: [13], [12], [11]

To measure the central tendency consider the mean, and the median - which in this case is simply the 50th percentile. Notice the fact that for both CCS and ICS the median and mean are nearly identical, for the CCS case it's less than one third of a second difference, and for the ICS case the difference between these two measure of central tendency is only slightly greater than a second. This points to both distribution being evenly distributed around the mean. As we can see for nearly all percentiles, minimum and maximum, and mean values the CCS times are roughly two thirds of the value of the ICS times, and all times for every category are significantly longer for the ICS sample relative to the CCS.

An indication of the relative spread can be observed by looking at the sample standard deviations. the ICS case is slightly more spread out than the CCS case $4.797s$ to $3.559s$ respectively, i.e slightly over a second more for the ICS sample. To gain a better descriptive understanding of the spread we can also look at the range, i.e. $R = x_{max} - x_{min}$ of the two samples:

$$R_{CCS} = x_{CCS,max} - x_{CCS,min} = 22.328000 - 8.630000 = 13.698$$
$$R_{ICS} = x_{ICS,max} - x_{ICS,min} = 35.255000 - 15.687000 = 19.568$$

From both the sample standard deviation and the range we can see that the spread of the ICS sample is greater both in terms of the range, i.e. outliers, and of the sample standard deviation:

Question 4: Plot the data

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.

Figure 2 points to the same ideas described in the descriptive statistics presented in the previous section. There are no real outliers for the CCS case, which shows up in the difference in the standard deviations of the two samples above, and that the mean and median for the CCS case are very close together. What is interesting to note, however is that the two greatest times for the ICS are from subjects who take no where near the longest for the CCS trial. Subject 20 having the second greatest ICS time isn't even to the right of the mean. This graph also explicitly shows that each subject takes a greater ICS time than their CCS time, there being no exceptions. One might think naively especially in

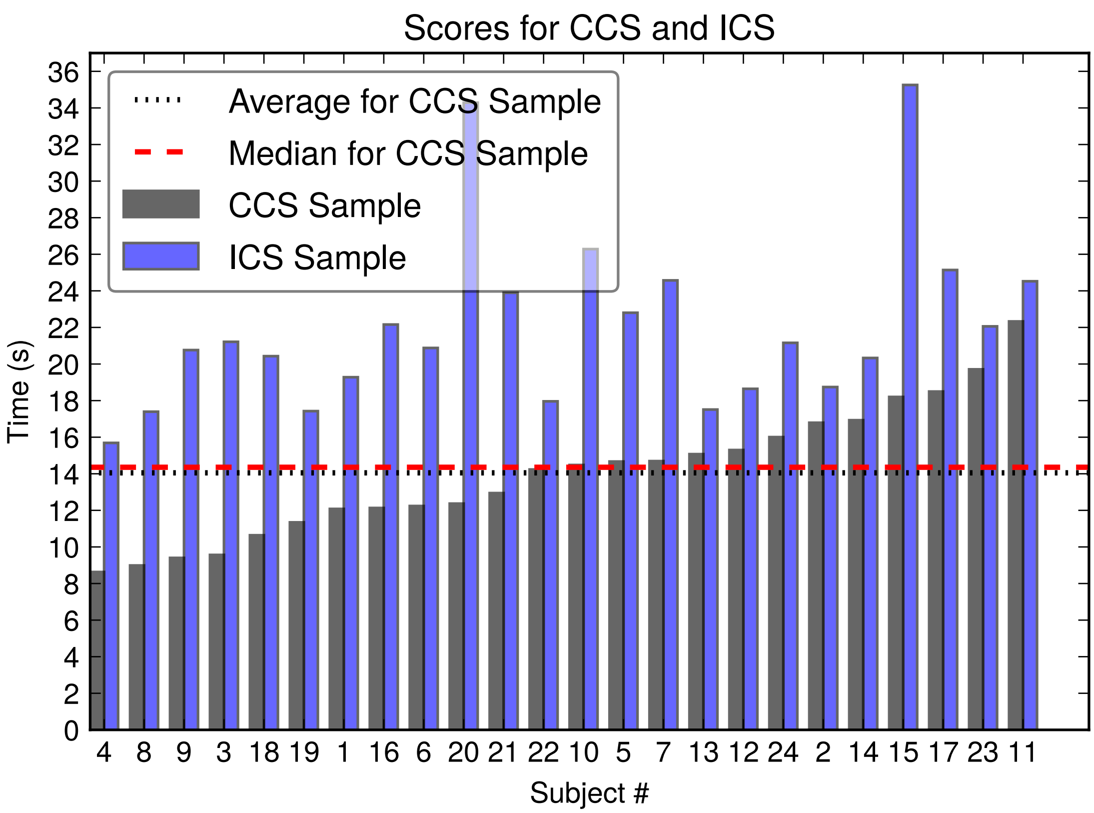

Figure 2: Bar graph of the time to complete the CCS and ICS list for each subject. The subject is labled in order of their position in the original data, however the data is ordered by CCS time, from least to greatest. The code for the plot can be found in p1_stroop_effect.py and demos and documentation from the official Matplotlib website were used to generate it, [9], [7], [14], [8], [10], [3]

the case of the CCS descriptive statistics that both data sets show an evenly distributed nature, however looking at the two histogram graphs, Figure 4 and Figure 3, it becomes very clear that both are heavily skewed to the left, with the two outliers on the right hand side for the ICS case increasing the standard deviation dramatically

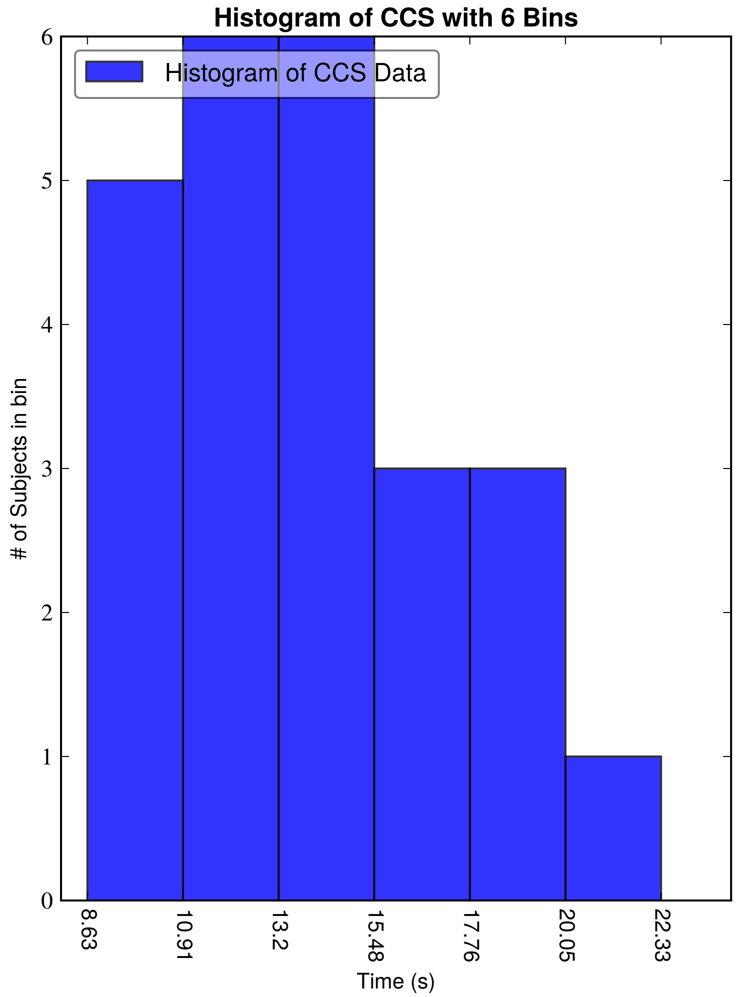

Figure 3: The graph above is a histogram of the CCS data. Note how it is skewed heavily to the left. This is due to the two outliers with times greater than $32s$.

As the sample size is not too great, it doesn't even exceed the minimum rule of thumb of thirty, and as such the shape of the population distribution isn't implied in the obviously left skewed sample distributions for both CCS and ICS, nor is the Central Limit Theorem to be relied upon given the sample size. So over generalizing is a bit too easy at this point, however even after removing the outliers in the ICS case, please see Figure 5, it is difficult to come to any definitvie conclusion about the distribution - only that the times are greater than the CCS data set.

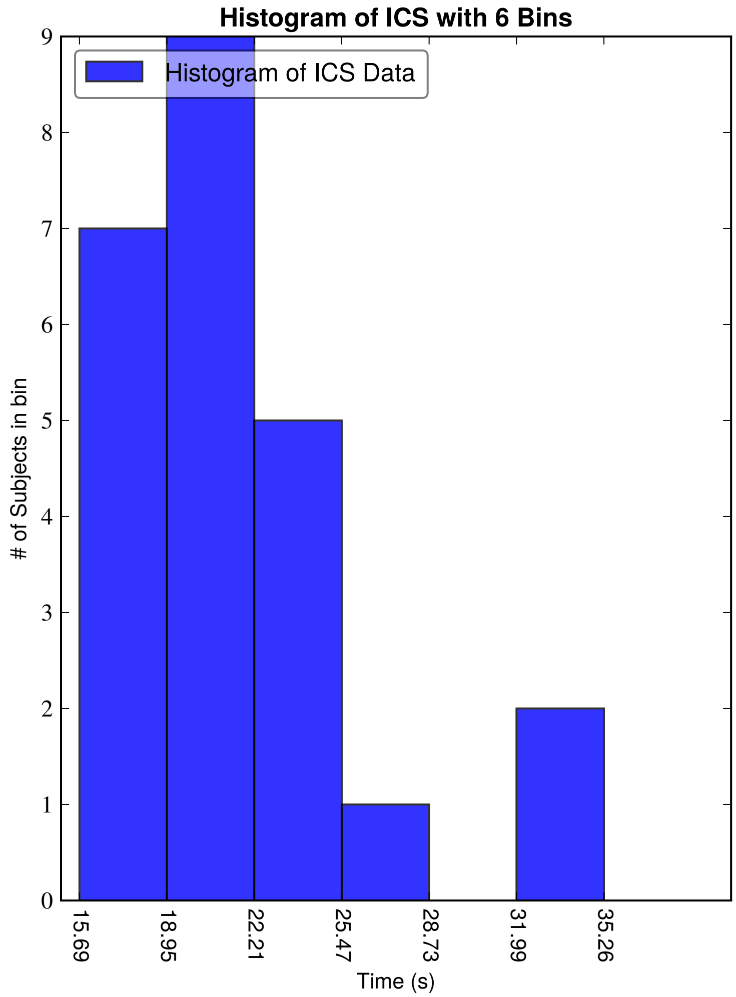

Figure 4: The graph above is a histogram of the ICS data. Note how it is skewed heavily to the left. This is due to the two outliers with times greater than $32s$.

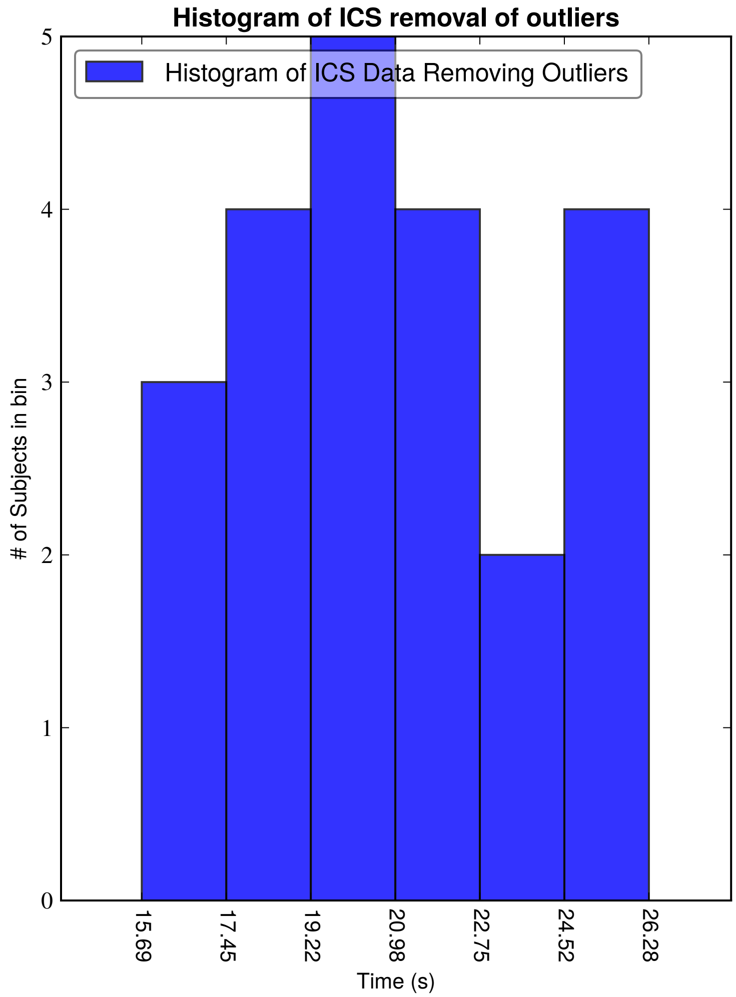

Figure 5: The graph above is a histogram of the ICS data where the two outliers whose times were greater than $32s$ have been removed.

Question 5: Perform the statistical test and interpret your results

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?

The test to be performed is a *t* Test for Correlated Groups. That is a test where the subjects of both groups remain the same and it's the independent variable CCS/ICS tests that distinguish one set of data from the other. For guidance concerning which test to apply I referenced the fine research methods book [6]. As stated in § 2 I believe both one and two sided tests can legitimately be argued for, however if this is a test regarding the Stroop effect only the one tailed test defined by the hypothesi discussed again in § 2 i.e. Equation 1 and Equation 2 are reasonable. Both § 3 and § 4 point to a statistically significant difference between the two data sets, and in this section the math is performed to show this, and quantatatively to what extent.

I have used the SciPy package to calculate the t value [1]. Below are the calculations as well as the results to this method:

```python
import pandas as pd
import numpy as np
import scipy as sp
import scipy.stats
from math import sqrt
stroopdata = pd.read_csv('stroopdata.csv')
print(scipy.stats.ttest_rel(stroopdata['Congruent'],stroopdata['Incongruent']))
(-8.020706944109957, 4.1030005857111781e-08)
```

Figure 6: Calculation of the two-tail correlated t test shows that the the null hypothesis should be rejected. That is $\mu_{ICS} > \mu_{CCS}$. For syntax highlighting, as I'm using pdfLaTeX, I used additional assistance from [18].

So to interpret the results. The first number obtained, -8.020706944109957 is the t value for the test and the second is the corresponding p value. That is the population mean, μ_{ICS} for the ICS data set is greater than the μ_{CCS} ($t = -8.020706944109957$ at $p = 4.1030005857111781e - 08$), an exceedingly unlikely occurrence, four in one-hundred million chance to occur if μ_{ICS} was in fact less than or equal to μ_{CCS}. Which means the test results points to H_A being the only acceptable hypothesis to accept, with almost nill chance of a Type I error.

The quantative results matched up well with the descriptive and inferential statistics discussed in sections § 3 and § 4. The degree to which the null hypothesis must be rejected after looking at the results of this test paint the stark difference between these two data sets.

Question 6: Digging deeper and extending the investigation

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!

There is a concise but equally inspired webpage that goes into the possible ways of viewing the Stroop effect, see [15]. What follows I cannot take credit for and must rely on the previously cited reference. They list three models explaining why the effect occurs: the automatic word recognition (AWR) model , speed of processing (SP) model, and a parallel distribued processing (PDP) model.

I will briefly describe these three models. The AWR is the claim that humans are hard wired for language and in this case word recognition. That is we are hard wired to process words in a certain manner, i.e. the meaning of a printed word takes prescedence over the color of the printed word. The evidence for this model is due to the robust nature of the Stroop effect. That the ICS sample always has a greater mean than the CCS sample, regardless of practice. That the use of color related words such as 'sky' or 'firetruck', to words that sound like color words such as 'wred' and 'bloo' all behave in accordance with the Stroop effect. So the AWR model claims that in essence the Stroop effect is due to how humans process information. That is we have a biological tendency to focus on, have a precedence for, semantic meaning over observable traits, like color.

The SP model makes the claim that certain neural pathways are simply faster than others. That the reading pathway is faster than the color processing pathway. As such it takes longer for the brain work out the conflicting messages and use the later arriving information regarding the color of the printed word over its semantic meaning. This is a poor example but I will use it anyways. Lightning to our eyes arrives faster than thunder to our ears. This is due to the nature of the medium that the energy is passing through. The metaphor breaks down quickly but it's an easy way of seeing the idea in practice.

Finally the PDP model claims that it's not so much a matter of speed as in the SP model but a matter of the strengths of the cognitive pathways of each. That is on the whole the pathway for reading is stronger and therefore takes presedence not because it arrives at the conscious decision making parts of our brain but because it makes more of an impact, so that in turn it requires more from us to focus in on the fainter observation of the color of the printed letters.

The Wikipedia article on this subject [4], is far more exhaustive in its description. It discusses the possible biological basis of the phenomenon, and provides good research references. Instead of parroting any further explication I will consider an experiment not in this article I thought would be interesting.

While originally I thought up the experiement on my own a bit of searching found that research had naturally already pushed into this area [2]. The only change to the normal Stroop effect would be to use experimental subjects all of whom had dyslexia, and a control of random but similarly aged readers. It seems to me that likely the dyslexic would have an easier time saying the name of the color of a certain printed word over reading the word which if it occurred would violate the Stroop effect. Studying the causes might hopefully shed light on why exactly the Stroop effect occurs, and provide more insight into dyslexia. Oddly what the researchers found in the referenced study is however just the opposite to the one I naively hypothesized. Namely the children with dyslexia had an even more pronounced Stroop effect than normal readers in their age group.

What is obvious is that the continuous advances in technology should help in understanding what the basis is for the Stroop effect. Another experiment is using a technology that isn't quite here but is a stone's throw away that of modern brain reading, or making correlations between uses of particular areas of the brain and our thoughts which in the last decade has really progressed immensely, see [5] as an example. Using this technology we would likely be able to see exactly where the extra time was coming from. Then again if real mind reading becomes a reality in any practical sense it likely would have far greater reaching effects but it's something to look forward to.

Statistics references

[2] C Faccioli et al. Poor readers but compelled to read: Stroop effects in developmental dyslexia. 2013. URL: http://www.ncbi.nlm.nih.gov/pubmed/17852135.

[4] Wikimedia Foundation. Stroop effect. 2015. URL: https://en.wikipedia.org/wiki/Stroop_effect.

[6] S. Jackson. Research Methods and Statistics: A Critical Thinking Approach. Cengage Learning, 2011. ISBN: 9781111346553. URL: https://books.google.com/books?id=YXHuw%5C_aIIgYC.

[15] Rochester Institute of Technology. BACKGROUND ON THE STROOP EFFECT. 2015. URL: https://www.rit.edu/cla/gssp400/sbackground.html.

Matplotlib references

[3] Python Software Foundation. 5. Data Structures. Sept. 2015. URL: https://docs.python.org/3/tutorial/datastructures.html#list-comprehensions.

[7] Matplotlib. colors: For a visual representation of the matplotlib colormaps, see the "Color" section in the gallery. Nov. 2015. URL: http://matplotlib.org/api/colors_api.html.

[8] Matplotlib. matplotlib.backends.backend_pdf. Nov. 2015. URL: http://matplotlib.org/api/backend_pdf_api.html.

[9] Matplotlib. pylab_examples example code: barchart_demo.py. matplotlib, Python. Nov. 2015. URL: http://matplotlib.org/examples/pylab_examples/barchart_demo.html.

[10] Matplotlib. pyplot: matplotlib.pyplot: Provides a MATLAB-like plotting framework. Nov. 2015. URL: http://matplotlib.org/api/pyplot_api.html.

[14] ssstuvz. Transparent legend for plots in matplotlib. Dec. 2013. URL: https://rykovanov.wordpress.com/2013/12/13/transparent-legend-for-plots-in-matplotlib.

SciPy references

[1] SciPy developers. scipy.stats.ttest_rel. 2015. URL: http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.stats.ttest_rel.html.

PANDAS references

[11] PANDAS. pandas.DataFrame.describe. PANDAS, Python. Nov. 2015. URL: http://pandas.pydata.org/pandas-docs/version/0.17.0/generated/pandas.DataFrame.describe.html.

[12] PANDAS. pandas.DataFrame.to_latex. PANDAS, Python. Nov. 2015. URL: http://pandas.pydata.org/pandas-docs/version/0.17.0/generated/pandas.DataFrame.to_latex.html.

[13] PANDAS. pandas.read_csv. Nov. 2015. URL: http://pandas.pydata.org/pandas-docs/stable/io.html%5C#io-read-csv-table.

Udacity references

[16] Udacity. Statistics: The Science of Decisions - Project Instructions. Nov. 2015. URL: https://docs.google.com/document/d/1-OkpZLjG_kX9J6LIQ5IltsqMzVWjh36QpnP2RYpVdPU/pub.

[17] Udacity. Statistics: The Science of Decisions Project Rubric. Nov. 2015. URL: https://docs.google.com/document/d/1-OkpZLjG_kX9J6LIQ5IltsqMzVWjh36QpnP2RYpVdPU/pub.

pdfLaTeX references

[18] Daniel Widerin. Syntax highlighting for Python scripts in LaTeX documents. 2013. URL: http://widerin.net/blog/syntax-highlighting-for-python-scripts-in-latex-documents.

other references

[5] Christian Herff et al. “Brain-to-text: Decoding spoken phrases from phone representations in the brain”. In: Frontiers in Neuroscience 9.217 (2015). ISSN: 1662-453X. DOI: 10.3389/fnins.2015.00217. URL: http://www.frontiersin.org/neural_technology/10.3389/fnins.2015.00217/abstract.