Name: Long Nguyen

Course: Udacity Data Analyst Nanodegree

Project: P1 – Test a perceptual phenomenon

1. What is our independent variable? What is our dependent variable?
   Our independent variable is the two conditions in which the participants have to perform the task. One is congruent words condition, in which the words are color words whose name match the colors in which they are printed. The other one is the incongruent condition, where the words are color words whose name do not match the colors in which they are printed.

   Our dependent variable is the time measured for each participant (or the mean time of all participants) when they perform the task in a particular condition. In specific, it's the time measured when a participant read the words in congruent condition and incongruent condition (or the mean time of all participants in each cases).

2. The null hypothesis for this task is that the mean time measured when everybody read congruent words is no different from the time measured when they read incongruent words. In other words, we are assuming that the average of the entire population of people who read the word in incongruent condition will be no different than the average of the entire population of people who read the word in congruent condition. The alternate hypothesis is that the incongruent words will slow down the reading time, which means that the mean reading time for anybody (not just people participating in the test) in incongruent condition will be more than that when he/she reads in congruent condition.

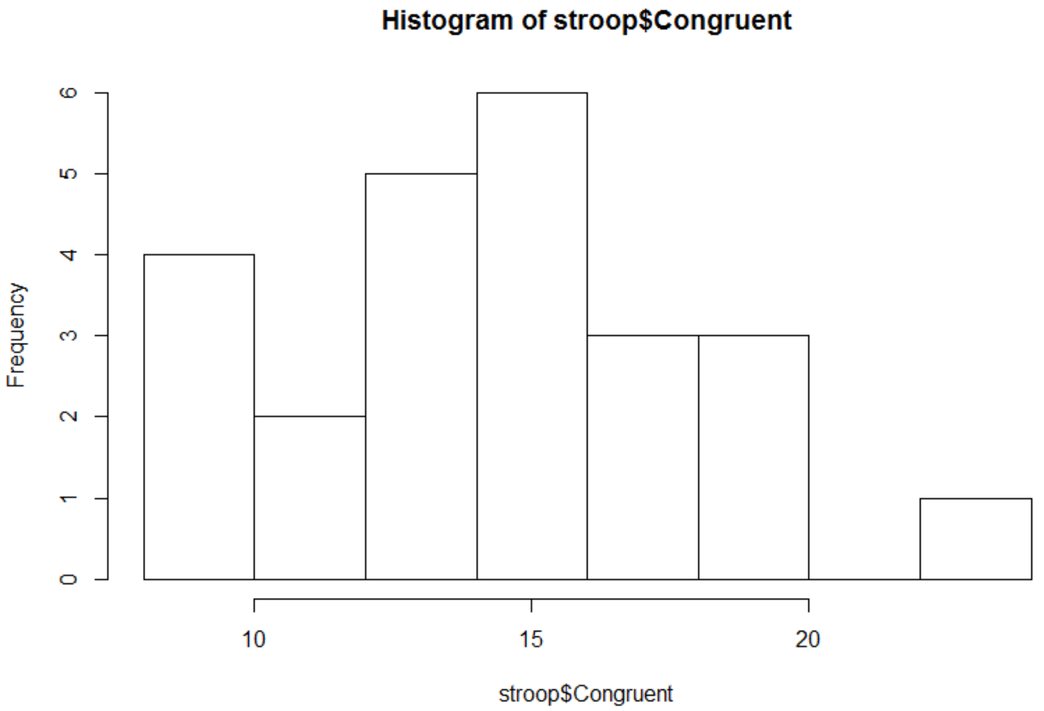
   In mathematical term:

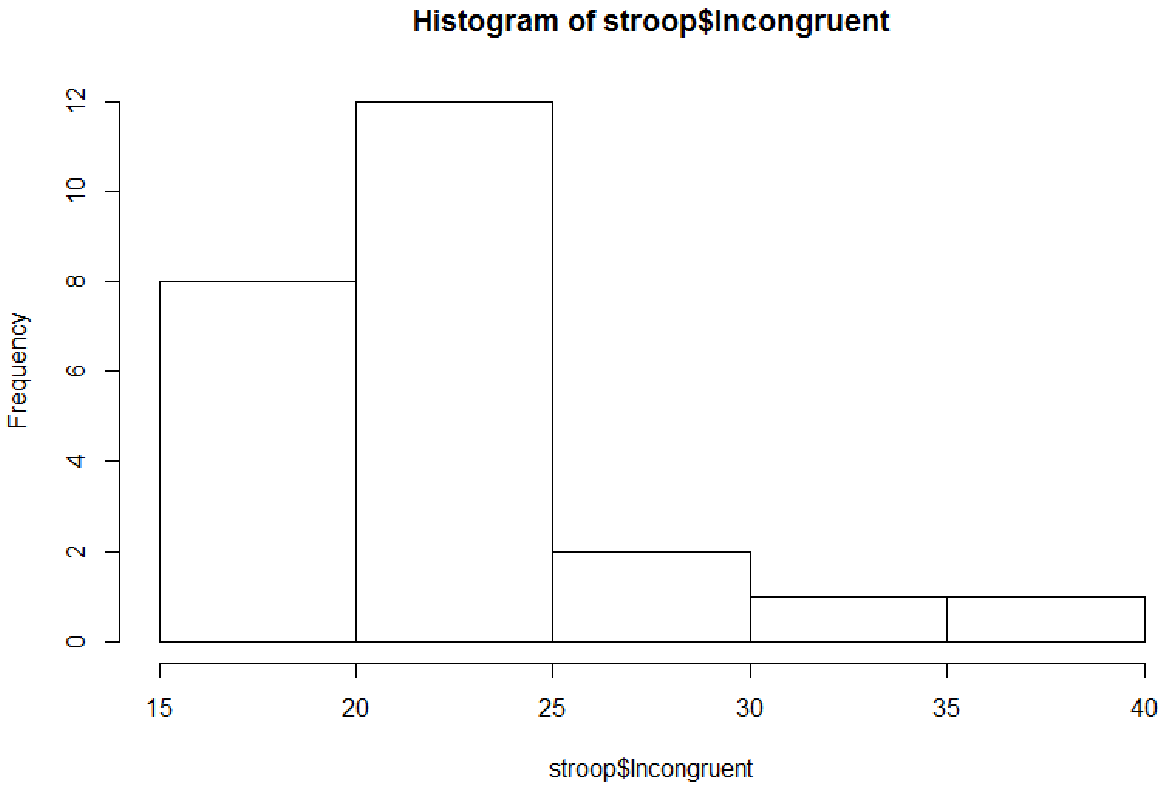
   Null hypothesis: $\mu_I - \mu_C = 0$
   Alternate hypothesis: $\mu_I - \mu_C > 0$

   The statistical test that I expect to perform on this task is the one-tail dependent t-test with $\alpha = 0.05$ (we use this α level because it's conventional). This test is chosen because the sample group is the same group and they receive two different treatment. In this test, the reading time of group A (congruent condition) and that of group B (incongruent) will be used to calculate the sampling distribution of sample means of two groups. These two distributions will then be used to create a hypothetical sampling distribution of the differences between sample means. Since population variances are unknown, we will estimate the standard error of this distribution using sample statistic: the combination of standard error of two sampling distributions of group A and B. Because we are using sample statistic, the distribution will not be normal and will follow a t-distribution. One-tail test is chosen because we are trying to see what is the probability that a particular differences will be bigger than 0,; we don't care if this difference is smaller than (we don't care if the incongruent words actually make people read faster).

3. The data is as follow:

Congruent	Incongruent
12.079	19.278
16.791	18.741
9.564	21.214
8.63	15.687
14.669	22.803
12.238	20.878
14.692	24.572
8.987	17.394
9.401	20.762
14.48	26.282
22.328	24.524
15.298	18.644
15.073	17.51
16.929	20.33
18.2	35.255
12.13	22.158
18.495	25.139
10.639	20.429
11.344	17.425
12.369	34.288
12.944	23.894
14.233	17.96
19.71	22.058
16.004	21.157


Histogram of stroop$Congruent
Frequency
0
1
2
3
4
5
6
10
15
20
stroop$Congruent


Histogram of stroop$Incongruent
Frequency
0
2
4
6
8
10
12
15
20
25
30
35
40
stroop$Incongruent

From the above histogram, we have these information about the data central tendency:

	Congruent	Incongruent
Mean	14.051125	22.01591667
Mode	14-16	20-25
Median	14.3565	21.0175

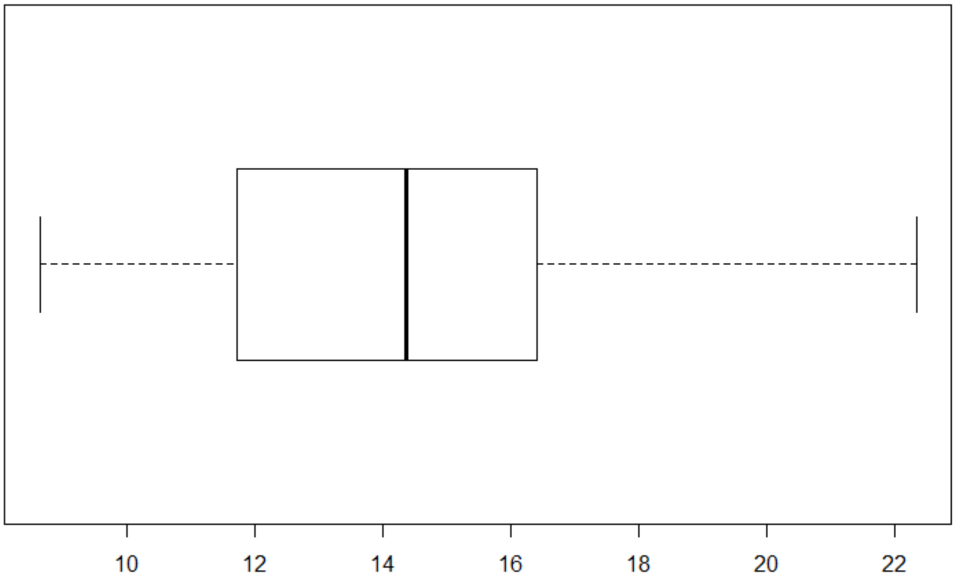
Table 1: Central tendency of congruent and incongruent

	Congruent	Incongruent
Estimated population standard deviation	3.559357958	4.797057122
1st quantile	11.52775	18.66825
3rd quantile	16.59425	24.3665
Inter-quantile range	5.0665	5.69825
Range	13.698	19.568

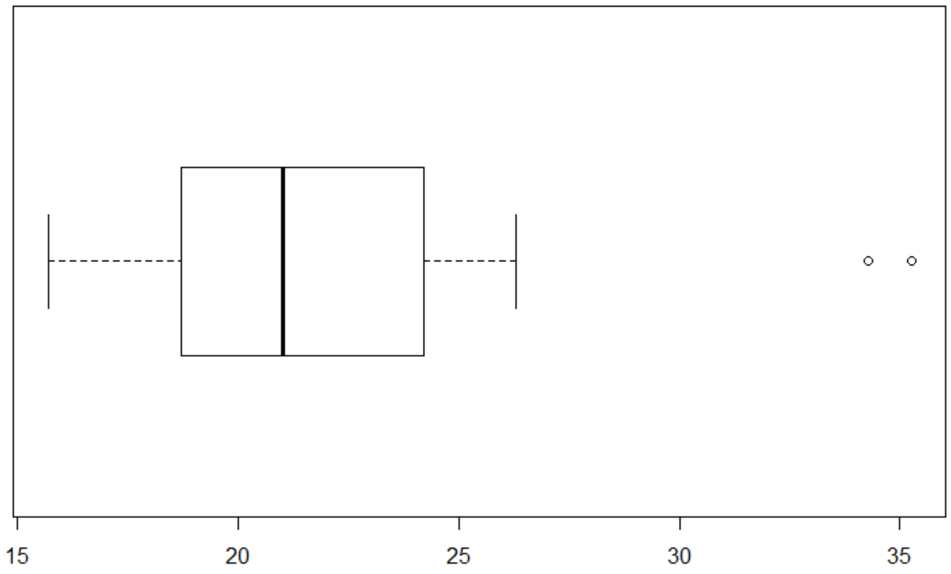
Table 2: Variability of congruent and incongruent

Box plot:

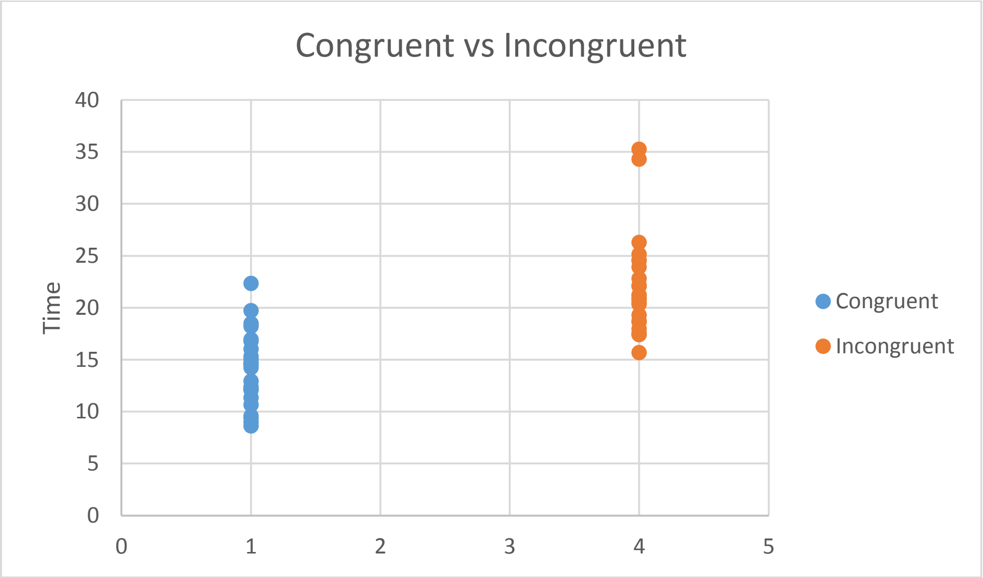
Congruent



Incongruent



4. Scatter plot of congruent vs incongruent:



As we can see from the graphs above, mean congruent is lower than mean incongruent. This is in line with what we measured in table 1 in section 3. From the histogram of congruent, we can see that the mode occurs in the range of 14-16. The histogram of incongruent, on the other hand, tells us that the mode occurs in the range of 20-25.

50 percent of the time measured in congruent condition centered in the range 12-16 while 50 percent of the time measured in incongruent condition centered around 19-25. This suggest that the data for incongruent will be more spread out than that of congruent. This is consistent with the information in table 2, which tells us that the within sample standard deviation and inter-quartile range of incongruent is slightly higher than that of congruent.

For the incongruent data, there are two outliers, which suggests that the mean of incongruent data might be overestimate.

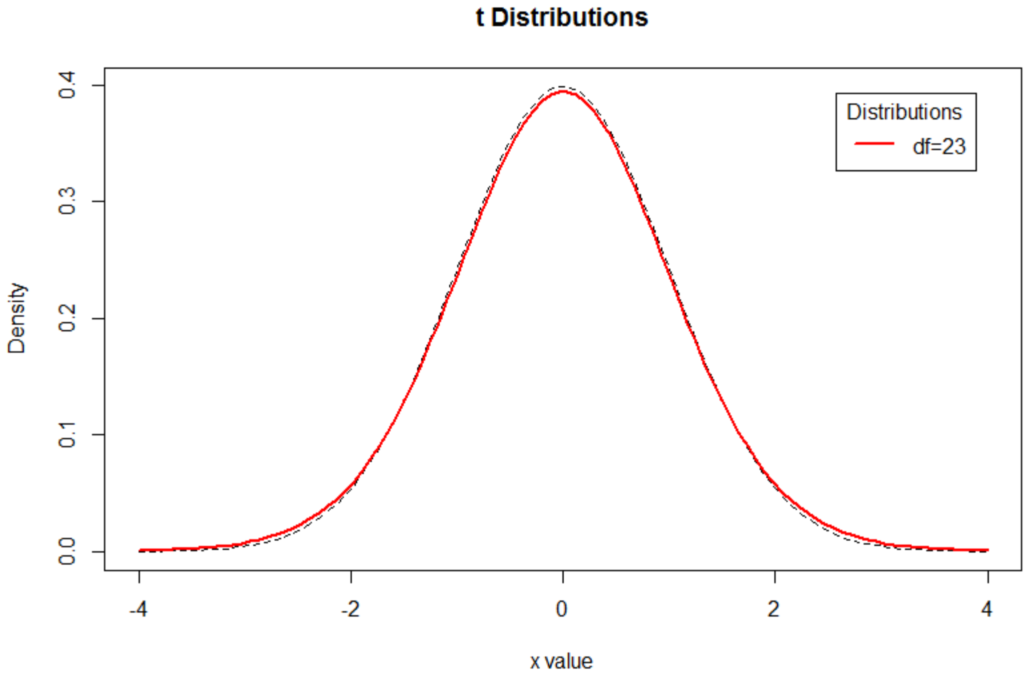
5. From the samples, we create another sample, which is the difference between the time measured in two samples:

Congruent	Incongruent	D = Incongruent - Congruent
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

The D sample belongs to a population N, with each observation is the time difference between congruent and incongruent reading condition. The sampling distribution of sample means of N will follow a t-distribution (since we have to estimate N variance using D statistics) with mean =

mean N = mean Incongruent – mean Congruent. The standard deviation of such distribution is: s_D/sqrt(n) = 0.9930

T distribution for 23 degrees of freedom is as follow:



Null hypothesis: $\mu_I - \mu_C = 0$ ($\mu_D = 0$)
Alternate hypothesis: : $\mu_I - \mu_C > 0$ ($\mu_D > 0$)

We will use the conventional confidence level of 95% to test our hypothesis. This makes our critical zone to be on the right tail with alpha = 0.05 and t-critical = 1.714

Assuming null hypothesis is true. From the sample we collected, mean Incongruent – mean Congruent = mean D = 7.965

If null hypothesis is true, the probability of selecting a mean difference of 7.965 form the sampling distribution of differences in sample mean is:

(7.965-0)/(s_D/sqrt(n)) = (7.965/0.9930) = 8.021 > 1.714

Because the chance of this happen by random is less than 5%, we can reject our null hypothesis. We can conclude that the mean of Incongruent is indeed larger than the mean of Congruent and this difference does not happen by chance.

The result matches my expectation that the time spent reading in incongruent condition will be significantly larger.

6. There are several theories that are used to explain this effect, and they are all known as “race model”. This is based on the assumption that both relevant and irrelevant information are processed in parallel in the brain, but “race” to enter the single central processor during response selection. The most common theory is the automaticity theory. It suggests that since recognizing colors is not an “automatic process” there is hesitancy to response, while the brain automatically understands the meaning of words as a result of habitual reading. (source: Wikipedia).

   An alternative task that would result in a similar effect would be the Numerical Stroop effect, which demonstrates the relationship between numerical values and physical sizes. A digit can be presented big or small, relevant or irrelevant to its numerical value. Results suggest that comparing digits in incongruent condition is much slower than in congruent condition.

References

Wikipedia. 18 9 2015. Article. 25 9 2015.

https://en.wikipedia.org/wiki/Stroop_effect