P1 - Test a Perceptual Phenomenon

Author: Ivailo Kassamakov, August 2015

Abstract: The present study is a statistical investigation of a perceptual phenomenon known as the Stroop Effect. This project is an effort towards fulfilling the requirements of the Data Analysis Nanodegree program offered by Udacity.

Note: The author used R Studio and LibreOffice Calc for performing the calculations and the diagram generation.

Introduction

In the present study, a descriptive and inferential statistical analysis will be performed on a dataset collected during subjecting a sample of randomly selected participants (n = 24) to a colored Stroop Test experiment.

The dataset consists of two series of time measurements – one done for the "congruent" task condition, the other one – for the "incongruent" condition (see below).

The goal is to perform the following analysis:

- identification of dependent and independent variables
- selection of appropriate hypothesis and statistical test
- presenting the descriptive statistics for the dataset
- providing appropriate data visualization and commenting on it
- performing the statistical test and deciding on hypothesis

The Stroop Effect

According to [wikipedia] the Stroop effect is a psychological phenomenon "of interference in the reaction time of a task". It can be easily demonstrated by performing a colored Stroop Test.

In its straightforward realization, the Stroop test consists of presenting the test subject with two sets of color-printed words. The words name colors.

In the first set (the "congruent" set) the color of each word matches the word meaning, i.e. the word "red" will be printed in red ink. For example: **RED**, **GREEN**, **BLUE**.

In the second set (the "incongruent" set), each word is printed in ink of color different than the denoted color. For example: **RED**, **GREEN**, **BLUE**.

Both sets are of equal size. The goal of the test subject is to name the ink color for each word in both sets. The time to name all colors in each set is measured.

The Stroop effect manifests itself in the generally longer times needed for naming the ink colors in the incongruent set.

The dataset

The dataset has been collected by measuring the times of several participants (n = 24), each of which has completed both parts of the the same Stroop test. Thus, the dataset contains 2 groups (for each of the congruent and incongruent sets) of 24 time measurements each.

The dataset is presented in Table 1.

Participant	Time, s	
	Congruent	**Incongruent**
1	12.079	19.278
2	16.791	18.741
3	9.564	21.214
4	8.63	15.687
5	14.669	22.803
6	12.238	20.878
7	14.692	24.572
8	8.987	17.394
9	9.401	20.762
10	14.48	26.282
11	22.328	24.524
12	15.298	18.644
13	15.073	17.51
14	16.929	20.33
15	18.2	35.255
16	12.13	22.158
17	18.495	25.139
18	10.639	20.429
19	11.344	17.425
20	12.369	34.288
21	12.944	23.894
22	14.233	17.96
23	19.71	22.058
24	16.004	21.157

Table 1: Original dataset for Stroop test with n = 24

The **independent variable** is the type of the set (congruent vs. incongruent). It is a categorical (nominal, factor) variable, having two levels: “congruent” and “incongruent”.

The **dependent variable** is the time for completing the set. It is a continuous-scale, ratio-type variable.

Choosing what to test

In an experiment like this we would be interested in proving whether there is a significant difference between the means of the times for completing the two experiment conditions.

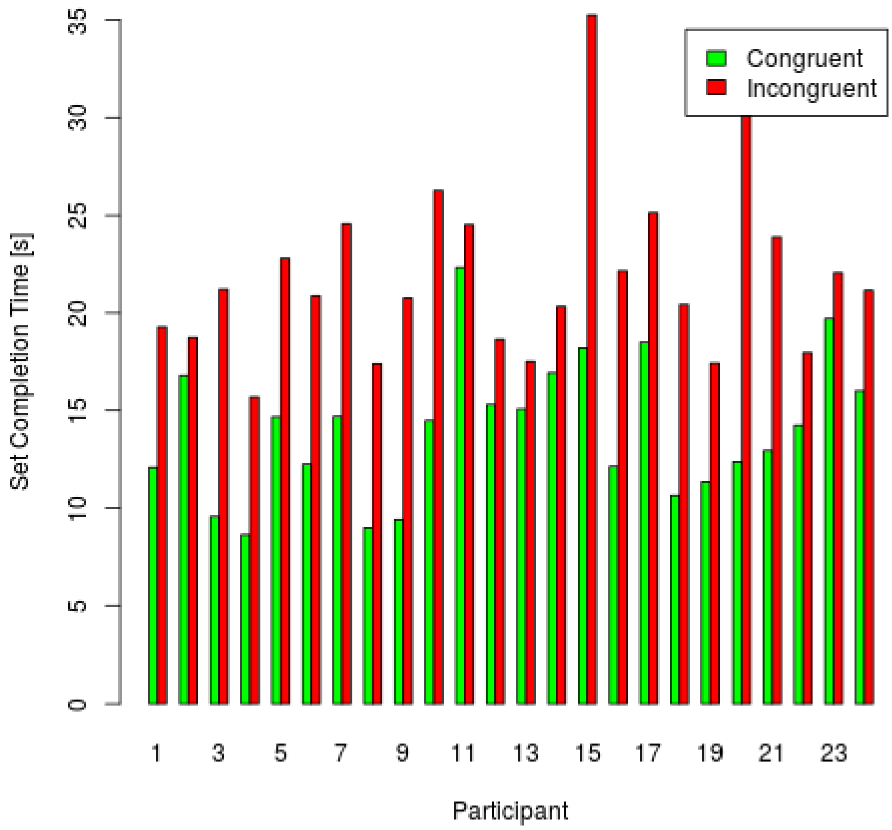
Since we have the same group of subjects performing both parts of the tests (i.e. we have a “two-conditions” **within-subject design**), we can perform our statistical testing by applying a **dependent-samples t-test**.

Before applying the test we will however need to make sure that the input dataset meets the following assumptions [Laerd]:

- The **dependent variable** is measured on a **continuous scale**
  - → OK, since our dependent variable is elapsed time measured in seconds
- The **independent variable** consists of two **categorical "related groups"**
  - → OK, since the same subjects are present in both set groups, and each subject has been measured on two occasions of the same independent variable
- The are no **significant outliers** in the **differences** between the related groups
  - → see below for outliers analysis
- The **differences** of the dependent variable between the related groups should be **approximately normally distributed**
  - → see below for normality test

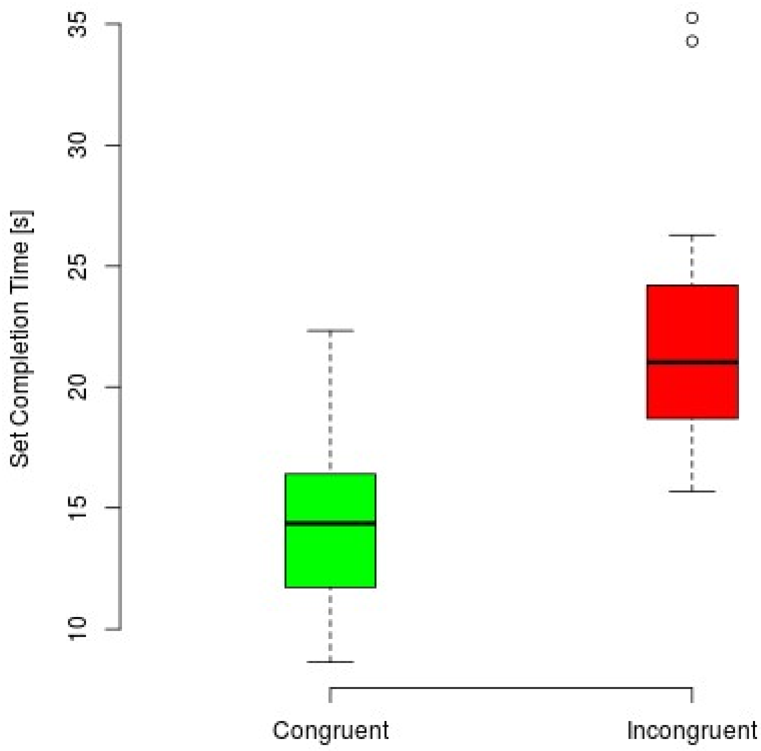
Exploratory data analysis of the two groups

Preliminary visualization



Plot 1: Bar plot for Stroop test with n = 24

We can gain a preliminary visual insight into the data via two commonly used plots: a bar plot (Plot 1) and a box-and-whiskers plot (Plot 2).



Plot 2: Box-and-whiskers plot for Stroop test with n = 24

Based on these plots we can draw the following visual conclusions:

- For 100% of the participants, the time for completing the incongruent set test is longer than the time for the congruent set.
- We have a strong indication that the median of the two measurement series are different.
- There are **two outliers** (> 1.5 * IQR) in the incongruent set measurements, as indicated by the points in the box-and-whiskers plot.

Outliers and data cleaning

As shown by the box-and-whiskers plot, there are two outliers in the incongruent set measurements. They correspond to cases #15 (t_{INC} = 35.255 s) and #20 (t_{INC} = 34.288 s).

Indeed, for the selected t-test, it is the **outliers of the between-group differences** that are important, not the outliers in the groups themselves. However, it is still useful to discuss and treat the outliers in the original dataset.

Of all causes for outliers listed in [Osbornet, et al] the most probable one in our case seems to be a "standardization failure". That is, these outliers can be caused, for example, by a momentary distraction of the subject during the test.

Although removing of outliers is normally not an easy decision, in our case removing them will not influence negatively the result of our test. The reason for this is the fact that the outliers are located at the upper range of the measurements, so they affect the incongruent group mean by increasing it.

Therefore, even if the outliers represent a valid data, by removing them we only make the statistical test for the alternative hypothesis more significant. If we manage to prove the alternative hypothesis without these outlying data values, then it will hold true also for the original dataset.

Consequently, we can go forward and safely clean the input data of these outliers by **pair-wise removing** the measurements for these two participants.

Participant	Time, s	
	Congruent	Incongruent
1	12.079	19.278
2	16.791	18.741
3	9.564	21.214
4	8.63	15.687
5	14.669	22.803
6	12.238	20.878
7	14.692	24.572
8	8.987	17.394
9	9.401	20.762
10	14.48	26.282
11	22.328	24.524
12	15.298	18.644
13	15.073	17.51
14	16.929	20.33
~~**15**~~	~~18.2~~	~~35.255~~
16	12.13	22.158
17	18.495	25.139
18	10.639	20.429
19	11.344	17.425
~~**20**~~	~~12.369~~	~~34.288~~
21	12.944	23.894
22	14.233	17.96
23	19.71	22.058
24	16.004	21.157

Table 2: Dataset (n = 22) cleaned of outliers

The following analysis will be performed with this outliers-free dataset with n = 22 (Table 2).

Descriptive statistics of the two groups

Table 3 shows some summary statistics of the cleaned dataset.

	Congruent	Incongruent
Mean, M	13.939	20.856
Standard Error, SE	0.766	0.613
Mode	none	none
Median	14.357	20.820
Variance	12.908	8.277
Sample Stdev, S	3.593	2.877
Kurtosis	-0.039	-0.777
Skewness	0.475	0.159
Range	13.698	10.595
Minimum	8.630	15.687
Maximum	22.328	26.282
Sum	306.658	458.839
Count	22	22

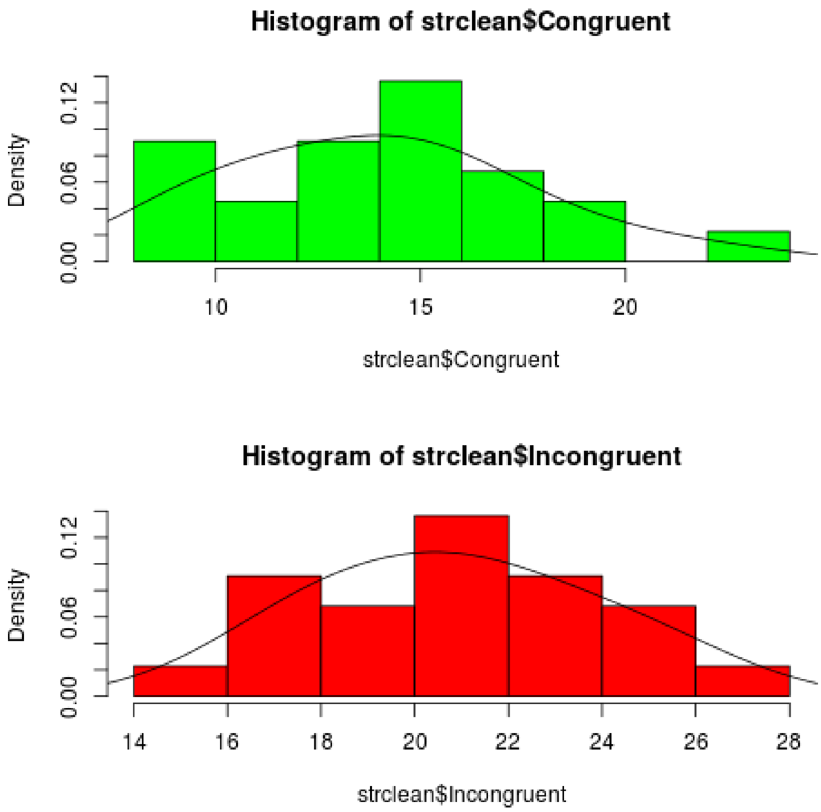
Table 3: Summary statistics for Stroop test with n = 22

Histograms, distribution and normality testing

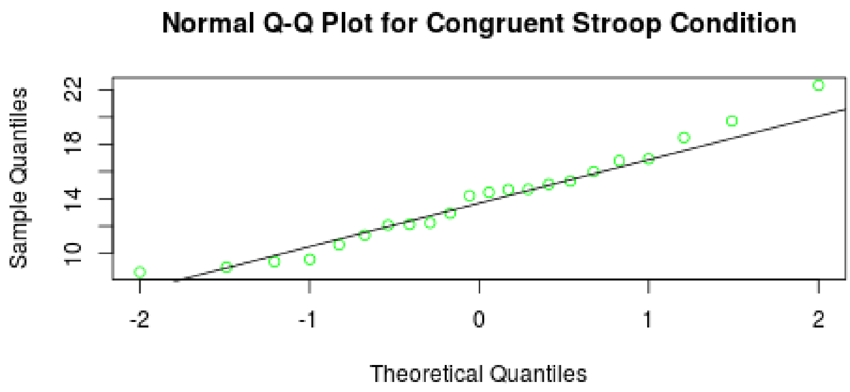
Next, we could explore the distribution of the measurement values in the two groups.

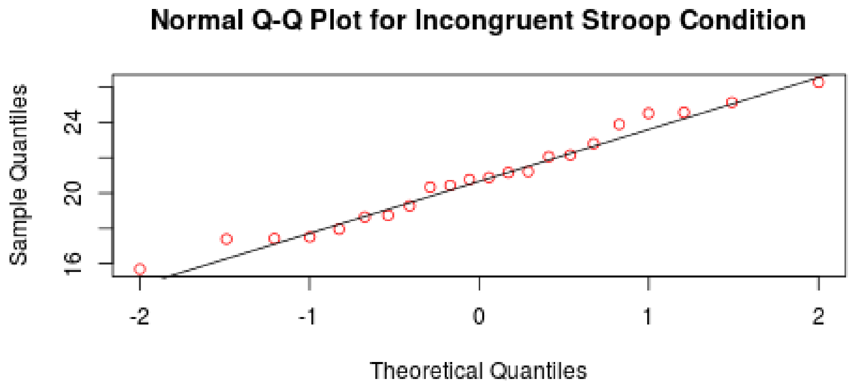
Plot 3 shows the histograms for both outliers-free measurement groups, with overlaid kernel density estimates. Visually analyzing it, we can hypothesize a **normal distribution**.

We can explore the normality hypothesis by utilizing a normal quantile (Q-Q) plot of the cleaned dataset (Illustration 4) - indeed, there is a strong indication for a normal distribution.



Plot 3: Histogram plots for Stroop test with n = 22





Plot 4: Normal Q-Q plots for Stroop test with n = 22

Finally, the Shapiro-Wilk normality test for both measurement groups returns the following values that strongly indicate a normal distribution:

Cleaned dataset, n = 22	**“Congruent”**	**“Incongruent”**
Shapiro-Wilk p-value	0.6182	0.7665

For comparison, we can also perform a Shapiro-Wilk test on the original dataset (with the outliers). The very small p-value for the incongruent group indicates convincingly that these outliers are most probably a noise on an otherwise pretty “normal” data:

Original dataset, n = 24	**“Congruent”**	**“Incongruent”**
Shapiro-Wilk p-value	0.6898	0.00259

We can conclude that the measurement data in both experiments is “pretty” (at α-level = 0.05) normally distributed.

Exploratory data analysis of the group differences

We now turn our attention to analyzing the pairwise differences between the two measurement groups, as these are the basis of the dependent-samples t-test.

Between-group differences

Table 4 shows the differences between the Incongruent and Congruent measurements for each participant of the cleaned dataset.

Participant	Congruent	Incongruent	DIFF(I-C)
1	12.079	19.278	**7.199**
2	16.791	18.741	**1.950**
3	9.564	21.214	**11.650**
4	8.630	15.687	**7.057**
5	14.669	22.803	**8.134**
6	12.238	20.878	**8.640**
7	14.692	24.572	**9.880**
8	8.987	17.394	**8.407**
9	9.401	20.762	**11.361**
10	14.480	26.282	**11.802**
11	22.328	24.524	**2.196**
12	15.298	18.644	**3.346**
13	15.073	17.510	**2.437**
14	16.929	20.330	**3.401**
15	removed	removed	**n/a**
16	12.130	22.158	**10.028**
17	18.495	25.139	**6.644**
18	10.639	20.429	**9.790**
19	11.344	17.425	**6.081**
20	removed	removed	**n/a**
21	12.944	23.894	**10.950**
22	14.233	17.960	**3.727**
23	19.710	22.058	**2.348**
24	16.004	21.157	**5.153**

Table 4: Between-group differences for the cleaned dataset (n = 22)

Descriptive statistics of the group differences

Table 5 shows some summary statistics of the between-group differences of the cleaned dataset.

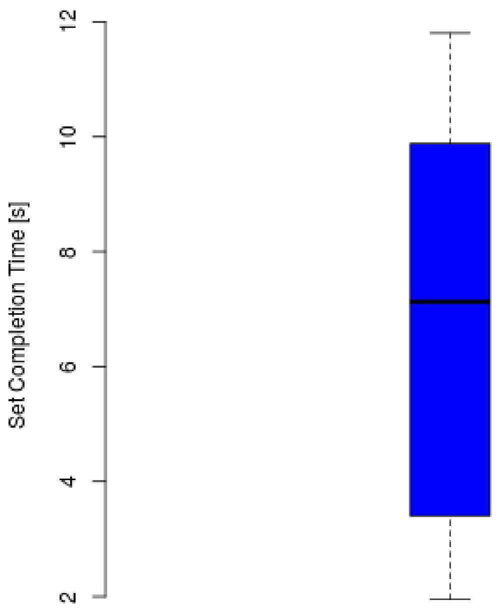
	DIFF(I-C)
Mean, Md	6.917
Standard Error, Sed	0.725
Mode	none
Median	7.128
Variance	11.564
Sample Stdev, Sd	3.401
Kurtosis	-1.417
Skewness	-0.100
Range	9.852
Minimum	1.950
Maximum	11.802
Sum	152.181
Count	22.000

Table 5: Summary statistics for the group differences (n = 22)

We see that, on average, the time for completing the incongruent set is longer by $6.917 \pm 3.401\,s$.

Outliers in the group differences

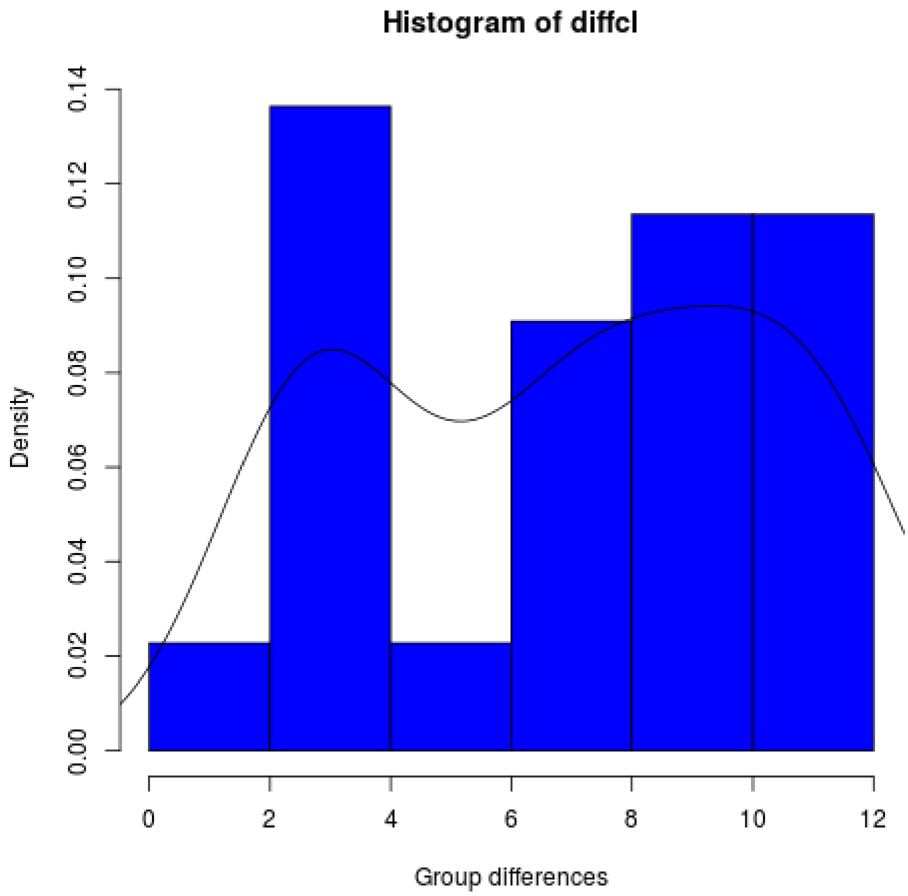
The boxplot of the group differences shows no outliers are present, fulfilling one of the requirements for the t-test.



Plot 5: Boxplot of the group differences for n = 22

Normality of the group differences

The histogram and the Shapiro-Wilk normality test show that the group differences do not exhibit normal distribution. Although not critical, this will make the application of a t-test less sensitive.



	Group differences, n = 22
Shapiro-Wilk p-value	0.074

Due to this deviation from normality, probably it would be more appropriate to use a non-parametric test, like Wilcoxon or McNemar. However, this is outside the scope of the current course, so we will go on as initially planned and apply a dependent-samples t-test.

Statistical hypothesis and t-test

As we saw before, there is a strong evidence that solving the incongruent Stroop tasks takes on average longer than the congruent task. We would like to **test statistically** whether the time for completing the incongruent condition test is significantly higher than the time for completing the congruent task.

To this end, we will perform a one-tailed t-test on the sample means for the two groups of measurements.

Hypothesis

The following statistical hypothesis can be formulated:

- **Null hypothesis**: There is no significant difference between the means of the time for the two parts of the experiment, or the incongruent time is smaller than the congruent one, i.e.:
  - $H_O: \ M_I \leq M_C$
- **Alternative hypothesis**: Solving the incongruent condition takes significantly longer than solving the congruent condition, i.e.
  - $H_A: \ M_I > M_C$ or, equivalently, $M_D = M_I - M_C > 0$

The null hypothesis will be tested via a **one-tailed t-test**.

t-test

We choose a standard α-level of 0.05. The degrees of freedom is $df = n - 1 = 22 - 1 = 21$. For one-tailed test we have a critical t-value $t^{crit}(21, 0.05) = 1.721$

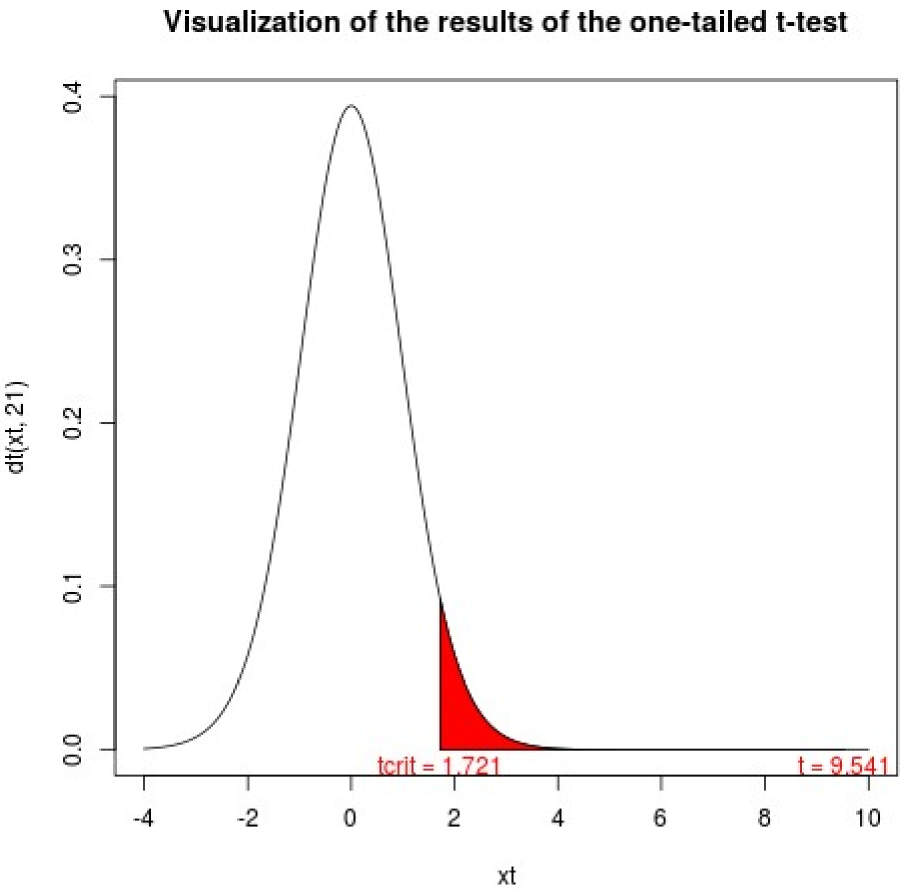
The t-statistic is as follows:

$$t = \frac{M_D}{S_D / \sqrt{n}} = \frac{6.917}{3.401/\sqrt{22}} = \frac{6.917}{0.725} = 9.541 \quad ,$$

where Md and Sd are the mean and the standard sample deviation of the group differences. The p-value for this statistic is 2.191e-09, which is practically 0.

Since the t-statistic is larger than the critical t-value we can **reject the null hypothesis**. We proved that solving the incongruent task takes significantly longer, t(21) = 9.541, p < .001, than the congruent task.

Plot 6 visualizes the results of the t-test.



Plot 6: Results of the one-tailed t-test

Confidence interval

The bounds of the 95% CI for the means' difference can be calculated as follows:

$$CI_{95} = M_D \pm ME = M_D \pm t(21, 0.05)\frac{S_D}{\sqrt{n}} = 6.917 \pm 1.721\frac{3.401}{\sqrt{22}}$$, where ME is the margin of error.

Therefore, $CI_{95} = [5.670; 8.165]$

Effect size

We can estimate the effect size via Cohen's d and the coefficient of determination r^2.

Cohen's d

$$d = \frac{t}{\sqrt{n}} = \frac{M_D}{S_D} = \frac{6.917}{3.401} = 2.034$$

This is very large (> 0.8) and shows that there is a consistent difference, on average, between the two experiment conditions.

Coefficient of determination

The coefficient of determination r^2 can be calculated from the t-statistic as follows:

$$r^2=\frac{t^2}{t^2+df}=\frac{9.541^2}{9.541^2+21}=.812$$

This means that 81% of the variance in the dependent variable is explained by the independent variable (the type of Stroop condition).

Discussion

The most plausible explanation of the observed differences in completing both parts of the Stroop experiments, is the so called “task interference”. In our case the interfering tasks is the cognition of the word meaning (i.e. reading and understanding the words), which interferes with, and consequently slows down, the main task of color recognition.

If the experiment is changed to include words in a language unknown to the respondent, the latter will complete both test parts in similar times. Thus, the interference is removed.

Other possible ways to eliminate the interference could be any graphical transformation (e.g. rotation or warping) of the words, that reduces the recognition of the words' meaning.

Many similar tests involving two simultaneous cognition functions can be designed. For example, counting the number of similar words. The “congruent” set in this case contains words that don't denote numbers, e.g. SUN, SUN, MOON, MOON, MOON. The “incongruent” set (exhibiting the interference effects) contains words that denote numbers, e.g. ONE,ONE, FOUR,FOUR,FOUR.

We could also think about combining a cognitive task, such as object counting, with the interfering task of perceiving a certain number of audible clicks.

Used bibliography

Interactive Stroop Effect Experiment

https://faculty.washington.edu/chudler/java/ready.html, Retrieved Aug 20, 2015.

Laerd Statistics - Dependent T-Test using SPSS
https://statistics.laerd.com/spss-tutorials/dependent-t-test-using-spss-statistics.php, Retrieved Aug 22, 2015.

Osborne, Jason W. & Amy Overbay (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research & Evaluation, 9(6).
http://PAREonline.net/getvn.asp?v=9&n=6, Retrieved August 24, 2015

NIST/SEMATECH e-Handbook of Statistical Methods,
http://www.itl.nist.gov/div898/handbook, Retrieved Aug 20, 2015.

Stroop Effect
https://en.wikipedia.org/wiki/Stroop_effect, Retrieved Aug 20, 2015.