# Udacity Data Analyst Nanodegree
# Project P1 – Test a Perceptual Phenomenon

Submitted by Narendran Santhanam

# Questions for Investigation

## 1. What is our independent variable? What is our dependent variable?

The independent variable is the incongruence of the words displayed, whereas the dependent variable is the time it takes to name the ink colors.

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

We are trying to find out if the incongruence of the words has a significant effect on the time it takes to identify the color of the ink displayed. Given below are some notations that we will use for this project:

- $\mu_C$: The population mean time taken to identify the ink color under the congruent condition
- $\mu_I$: The population mean time taken to identify the ink color under the incongruent condition
- $\bar{x}_C$: The sample mean time taken to identify the ink color under the congruent condition
- $\bar{x}_I$: The sample mean time taken to identify the ink color under the congruent condition

The null hypothesis for this experiment is that the time taken to identify the color does not significantly vary between the two conditions. The alternative hypothesis is that the time taken does vary between the two conditions.

$$H_0: \mu_C = \mu_I$$

$$H_A: \mu_C \neq \mu_I$$

A two tailed T-test would be appropriate for this experiment for the following reasons:

- We do not know what the population mean and variance are. Thus, a z-test cannot be performed to compare the sample mean with the population mean.
- The sample size is too small (<30) and the t-test is more suitable under such conditions.
- This is a dependent samples experiment, with each subject taking the same test twice under different conditions.
- We do not know if the population mean time taken for the incongruent condition will be greater than or less than the population mean time taken for the congruent condition, thus, a two-tailed test is chosen.

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

Here are some statistics derived from the dataset:
1. Sample size, $n = 24$
2. Sample mean time taken under the congruent condition: $\bar{x}_C = 14.051125$
3. Sample mean time taken under the incongruent condition: $\bar{x}_I = 22.01591667$
4. Point estimate for $\mu_C - \mu_I = -7.964791667$
5. Standard Deviation of the difference between the two samples: $S = 4.86482691$

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.

Here is a plot of the time taken to identify the ink color for all the participants for both the conditions:
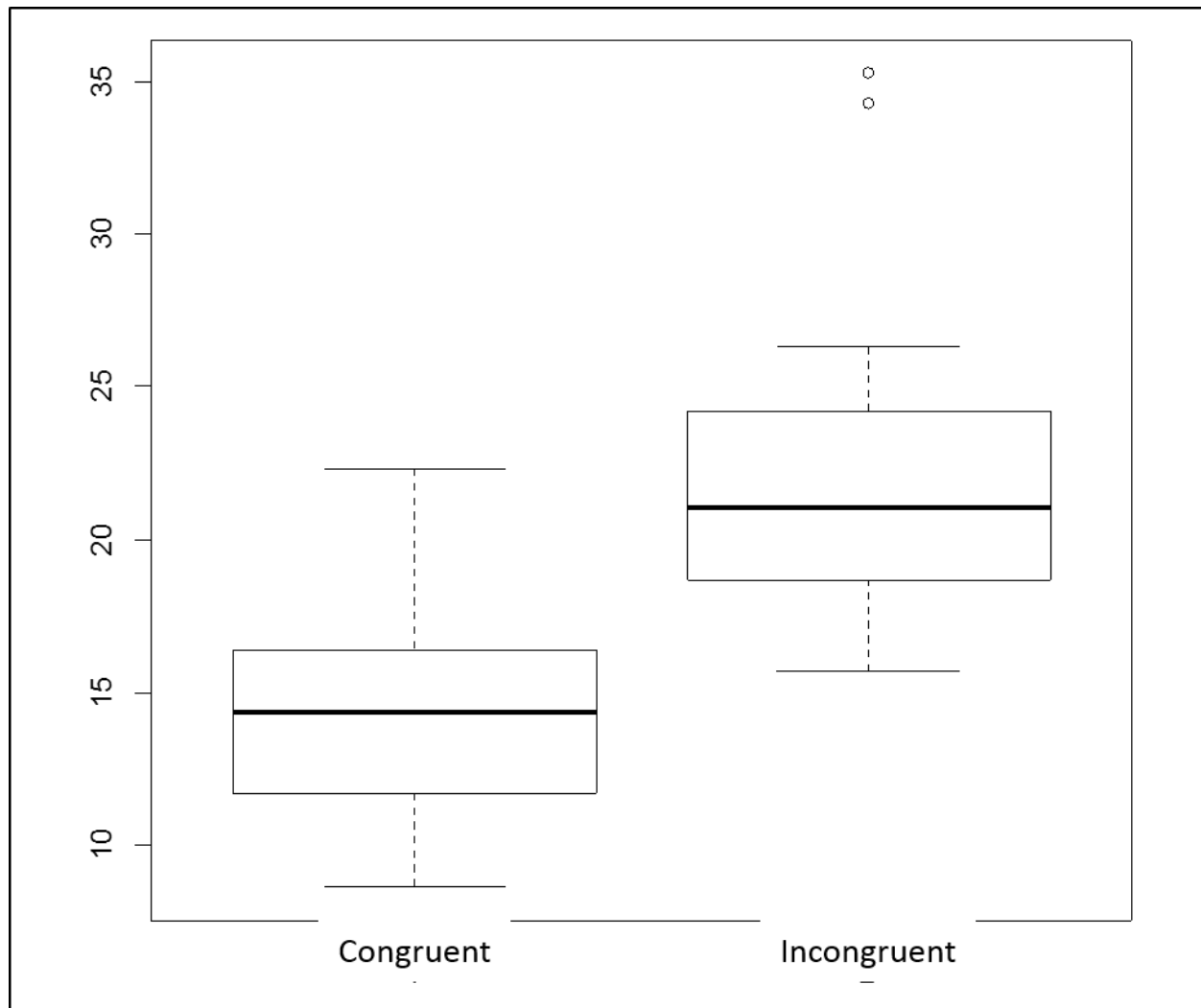


- We can see that the time taken for the incongruent condition is well ahead of the time taken for the congruent condition. Whether the difference is significant enough to conclude that incongruence does affect the time taken to identify the ink color is something that we can ascertain after performing the statistical test.

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?

Let us calculate the t-statistic for the given data.

$$t = \frac{\mu_C - \mu_I}{S/\sqrt{n}} = \frac{-7.964791667}{4.86482691/\sqrt{24}} = -8.020706944$$

Let us find out what is the t-critical for this experiment at 95% confidence level ($\alpha = 0.05$). This being a two-tailed test, we have to look for the t-critical value for p=0.025 and 23 degrees of freedom (n=24). The t-critical values obtained are **±2.069**.

We can see that the t-statistic lies in the critical region. Thus, we reject the null hypothesis and conclude that the incongruence between the words and color of the words displayed does have a positive effect on the time taken to identify the color of the ink. This result matches our expectations earlier from question 4 where we saw that the time taken for the incongruent condition was consistently higher than that for the congruent condition.

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!

I think the effect is due to the fact that text gets processed in real time in our brain, whereas identifying colors probably happens in the backend. For example, when I took the test, I found that in the incongruent condition, subconsciously my brain was searching for the color against a previously stored "table" of colors if you will, and came out with the answer. In the congruent condition, I was just reading out the words as I saw them. The Wikipedia page for Stroop effect lists four theories that could explain this phenomenon.

I think a similar effect may be achieved using direction as the parameter (instead of color). For example, on a map of a country, if we display words like "North", "South", "East" and "West" in different corners of the map we can achieve congruence and incongruence by asking participants to identify the direction in which the word is displayed. See an illustration below:
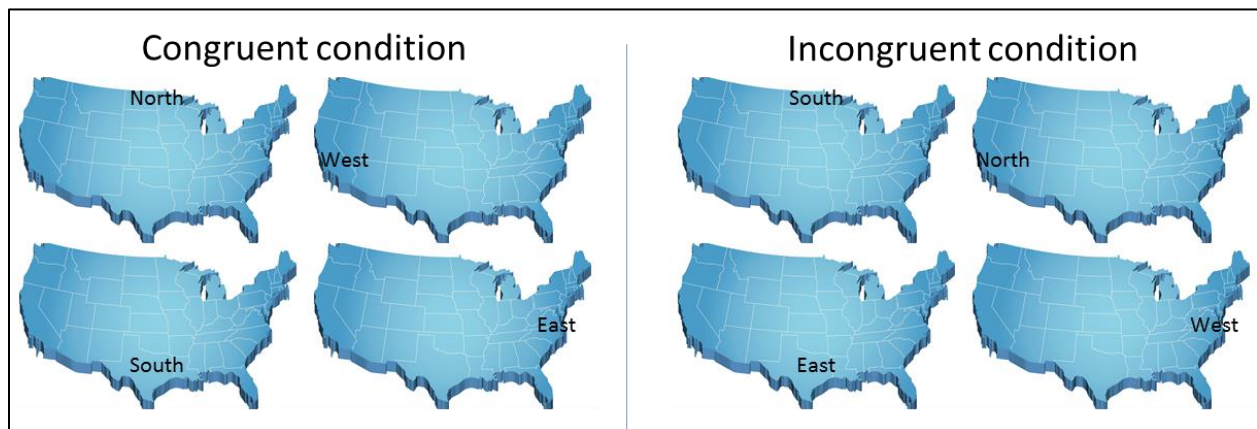


References:

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