PROJECT: TEST A PERCEPTUAL PHENOMENON

QUESTIONS FOR INVESTIGATION:

Q1: What is our independent variable? What is our dependent variable?

1. The independent variable in this project is **word set (congruent/incongruent).**
2. The dependent variable in this project is **time taken to name the ink colours of words.**

Q2: What is an appropriate set of hypotheses for this task? What kind of statistical test do you expect to perform? Justify your choices.

1. The experiment consists of creating a set of words for two conditions: congruent words and incongruent words. Then, time taken to name the colour of these words is measured for each condition. This is a type of **Repeated Measures Test (two conditions)** problem. We can think the variable as **time taken to name the ink colours of words** as mentioned earlier. Firstly, measure the variable for one condition (for congruent words) and then for another condition (making the set incongruent words), measure the same variable.
2. Based on the method above mentioned, following set of hypotheses is constructed:

$$H_0: \mu_d = 0 \quad \text{(Null Hypothesis)}$$
$$H_A: \mu_d \neq 0 \quad \text{(Alternative Hypothesis)}$$

Where: $\mu_d = \mu_2 - \mu_1$ (difference between means of 2 conditions)
μ_1 = Population time taken mean for congruent word set
μ_2 = Population time taken mean for incongruent word set

The methodology the hypothesis is basically to determine statistically, whether the values of these two conditions for the paired samples are significantly different from each other. From the above set of hypothesis equations, it is clear that we are performing **bidirectional two tailed** test. Following are the two statements formed from the hypothesis:

H_0: Time taken by people to name ink colour of word remains same for two conditions.
H_A: Time taken by people to name ink colour of word in the incongruent set is different than one required in the congruent set.

3. After performing analysis, we can determine which statement is true. We will be performing statistical test: **t-test.**

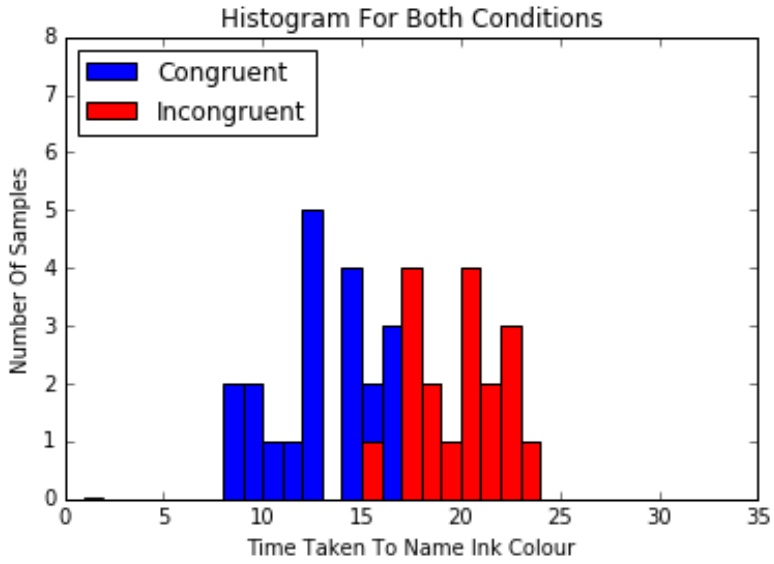
4. **General Assumptions:**
    - Observations are independent of one another. That means that the people under taking the experiment are not related to each other.
    - The dependent variable (time taken) is approximately normally distributed.
    - The dependent variable does not contain any outliers.
5. **Motivation:**
    - Repeated measures is selected instead of independent groups because we want to observe the results under minimum variations. In the independent groups case, the people in the two conditions will be different can hence it can introduce variations in the observations due to individuality.
    - T-test is selected instead of Z-test because the **sample size is below 30** and the **standard deviation of the population is unknown.** These two are the most important thumb rules to consider while selection of the statistics test.
6. **Assumptions for t-test:**
    - Scale of measurement for the dependent variable follows a continuous/ordinal nature.
    - Data collected for the sample is randomly selected from the population pool.
    - The collected sample data when plotted, results in a normal/bell-shaped curve distribution.
    - Homogeneity of variance. This means that equal variance exists when the standard deviations of samples are approximately equal.

Q3: Report some descriptive statistics regarding this dataset. Include at least one measure of central tendency and at least one measure of variability.

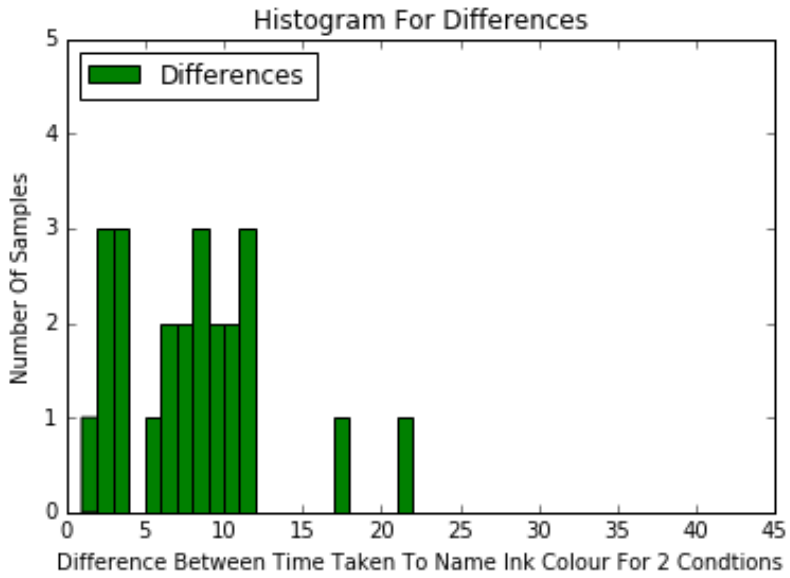
1. Measure of central tendency: **Sample mean of the differences (X_D)**
   The value obtained for X_D is 7.96. It is the mean value of the differences between the two conditions. Calculation is done in excel sheet under the name **Average D**. This value must be approximately normally distributed.
2. The means for both the conditions is also calculated. They are calculated as **Mean1** = 14.05 & **Mean2** = 22.01 in the excel sheet.
3. Measure of Variability: **Standard Deviation (S_D)**
   The sample standard deviation of the differences S_D is equal to 4.86. When the population size is more than 10 times larger than the sample size, then the standard deviation can be approximated by dividing the S_D value by square root of sample size. This value is equal to 0.99 **(standard error)**.
4. Measure of effect size: **Cohen's d**
   This factor indicates the extent of effect size between two conditions. Dividing the sample mean of the differences by its standard deviation **(X_D/S_D)** gives Cohen's d. The value obtained for Cohen's is 1.63. According to standard interpretations, any value above 1 implies **a large effect.**
5. Measure of effect size: **R^2**
   This factor explains the same thing as mentioned above for Cohen's d, but by a different way. This is given by dividing T_Statistics^2 by T_Statistics^2+DOF **(t^2/(t^2+DF)).** The value of R^2 is equal to 0.73. According to standard interpretations, any value above 0.45 implies **a large effect.**

Q4: 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.

1. In the first plot, a histogram for dependent variable of both conditions is plotted. In earlier point, we had made an assumption that the dependent variable must be approximately normally distributed. From this plot, we can say that assumption is reasonable.



2. The second plot is for the histogram of the differences between two conditions. (dependent variable) The nature of this data is also approximately normally distributed. This data is very important as the statistics t-test is performed on it.



3. Code for these two plots is included in the python file named: 'P1_Visual.py'

Q5: 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?

1. Statistical test is performed step by step in following manner. Firstly, difference between two condition values is obtained. Sample mean of the differences (X_D) is calculated equal to 7.96.
2. Standard deviation for these differences is computed, which is equal to 4.86. Standard error is obtained by dividing S_D by square root of 24. (sample size) This is equal to 0.99.
3. T_Statistics is computed by dividing X_D by Standard error. The value is equal to **8.02.**
4. Next, the critical statistic value (T_C) is extracted from the t-table. We look in the column of 0.025 (This value is taken for 2 tailed test. Alpha is 0.05) & in the row of 23 (Degree of freedom). We get the value of T_C equal to **+/- 2.069**.
5. The 95% confidence interval (CI) is given by: (X_D +/- T_C*Standard error) This 95% CI value is calculated as: **(5.91,10.01)** with margin of error (MOE) equal to **2.05.**
6. The p-value obtained from 'graphpad' website is: **p < 0.0001.** This value means the probability for observing the results under null hypothesis are less than 0.0001.
7. By the above arrived results, it is concluded that: **We have to reject the null hypothesis.** It can be interpreted that the change in the time taken to name the ink colour of the words is **statistically significant.** (more in this case as it is in the upper tail) Hence, we can select the alternative hypothesis that time taken to name ink colour for incongruent word set is different than required for congruent word set. (more time is required)
8. Another conclusion can be made that while the true mean change (or increase in this case) in the time taken to name ink colour of the words is 5.91 to 10.01 seconds. (We are 95% confident about this interval)
9. My expectation for this experiment was to **accept the null hypothesis.** But the results obtained were counterintuitive in nature as the change in the variable was extremely significant. (statistically speaking)

Q6: 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!

1. **Explanation:** The brain processes the information we receive from our sensory organs. (eyes in this case) First, we see the word given to us. Then the brain processes the word and obtains all the information related to that word from the memory. The colour name associated with the word is also acquired in that information. This is our **word processing** process. Also, the brain gets the colour of the word from our sensory organ eye when we look at the word. The brain then looks in our memory to find the name of the colour it has received from the eye. This process is called **colour identification**. In the congruent word set, the word and the name of word ink is same. So, the brain arrives at the conclusion by word processing process only. But, in the incongruent word set, the brain first arrives at the name of the word colour by word processing process. This colour name is different than the actual result. The brain also receives the name of the word colour by the colour identification process (which takes longer), which is correct colour name in this case. The fact that word processing is faster than colour identification is because of the evolution.
2. **Similar effect**: Stroop effect can be observed when we conduct a similar test using word/shape combination. In one set we create combination of shape (like circle, square) and its name in the shape. (congruent set) While in another set (incongruent set), a combination of shape and a name different shape name. (like word 'square' inside a circle) The time required to name the shapes will be more for incongruent set than congruent set.

REFERENCES:

Stroop effect Wikipedia page