Udacity Data-Analyst Nanodegree
P1: Statistics: The Science of Decisions Project Instructions

Project Link
https://docs.google.com/document/d/1-OkpZLjG_kX9J6LIQ5IltsqMzVWjh36QpnP2RYpVdPU/pub?embedded=True

Questions For Investigation

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

A. Independent Variable: Variables which are potential cause of variation in the sample output.
Conditions of tasks: a congruent word condition and an incongruent word condition
Dependent Variable: Variables whose variation in value is being studied.
Time taken to name the ink colors in equally sized lists

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

A.
A hypothesis test examines two opposing hypotheses about a population: the null hypothesis and the alternative hypothesis.
Null hypothesis (H0)
The null hypothesis states that a population parameter is equal to a value which means in our case the two sample are provided who belong to the same population.
Alternative Hypothesis (H1)
The alternative hypothesis states that the population parameter is different from the value of the population parameter in the null hypothesis which in our case is that the two samples do not belong to same population.

Comparing the two sample population, following would be the appropriate set of hypothesis:

Set of Hypothesis:

Null Hypothesis: The average population time taken to name the ink colors in congruent word condition task is **not** statistically different from the average population time taken to name the ink colors in incongruent word condition task.

Alternative Hypothesis: The average population time taken to name the ink colors in congruent word condition task **is** statistically different from the average population time taken to name the ink colors in incongruent word condition task.

$$H_0: \mu_{congruent} \neq \mu_{incongruent}$$
$$H_A: \mu_{congruent} = \mu_{incongruent}$$

where,
H_0 is the Null Hypothesis;
H_A is the alternative hypothesis;
$\mu_{congruent}$ is the population mean time taken to name the ink colors in congruent word condition task; and
$\mu_{incongruent}$ is the population mean time taken to name the ink colors in congruent word condition task;

Statistical Test: We will be performing a **Two-Tailed dependent t-test** on both the samples. The experiment scenario is a **dependent two-conditional** (the two conditional tasks) experiment which deals with within-subject design. Also since we are unaware of the population parameters(mean, standard deviation), we choose to perform the dependent t-test.

We compare the performance of both conditional task using t-test. We calculate the *t-statistical* value which indicates how many standard deviation away the observed value will lie from sample mean. We then find out the *t-critical* value based on the alpha level that we choose(using t-table) and identify the *t-critical* region. If the *t-statistical* value lies in the *t-critical* region, we can reject the Null Hypothesis. However, if the *t-statistical* value does not lie in (or is outside) the *t-critical* region, we fail to reject the Null Hypothesis and thus retain the Null Hypothesis. Rejecting Null Hypothesis signifies that the samples results are significantly different from each other, and Retaining the Null Hypothesis signifies that the sample results are not significantly different from each other.

Type of Test: This will be a **two-tailed test** as we do not have a direction in the hypothesis. Thus the two samples can be significant in both the directions (positive or negative).

Taking the absolute value in to account, mathematically, if $|t_{statistical}| > |t_{critical}|$ we reject the Null Hypothesis.

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

Measure of central tendency		
	Congruent	**Incongruent**
Mean	14.05	22.02
Median	14.36	21.02

Measure of variability		
	Congruent	**Incongruent**
IQR	-4.194	-3.798
Range	13.70	19.57
Variance	12.67	23.01
Standard Deviation	3.56	4.80

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.
A. We can visualize the data with the help of visualization tools available in google sheets.

Figure 1 and 2 shows a histogram chart for the data of Congruent and Incongruent Conditional Task. We can clearly see that the both the distributions are normal. Figure 1 is very clearly uniform with most of the sample population taking 12-15 seconds to perform the task. According to the data, most of the sample population takes 20-24 seconds to perform Incongruent Conditional Task.

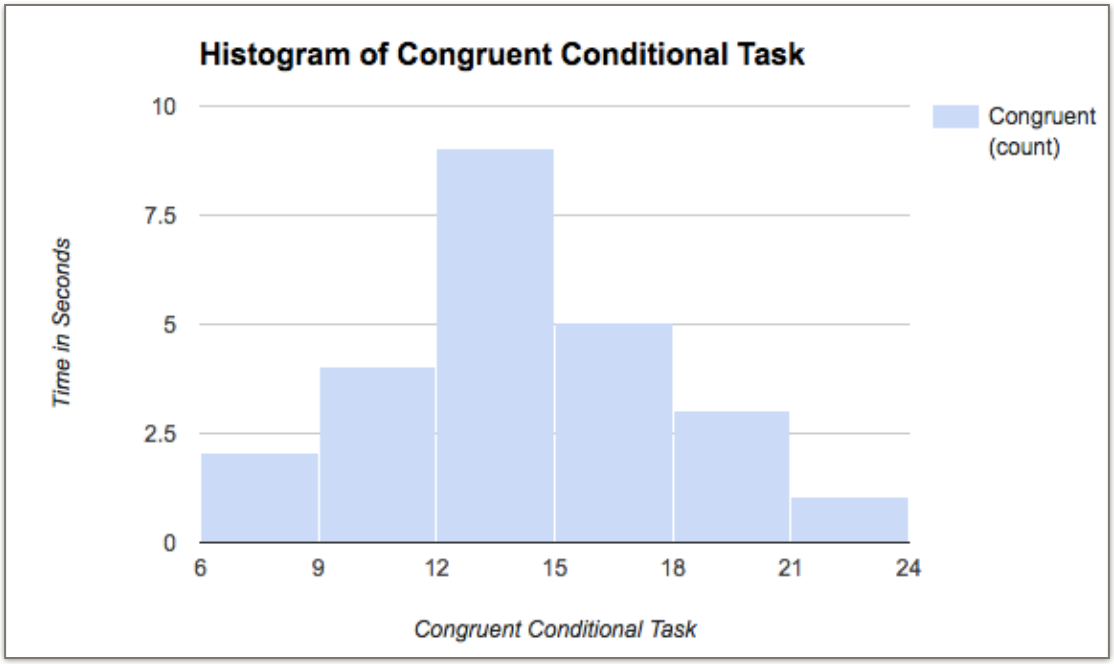


Figure 1: Histogram showing the time taken for Congruent Conditional Task with bin size 3

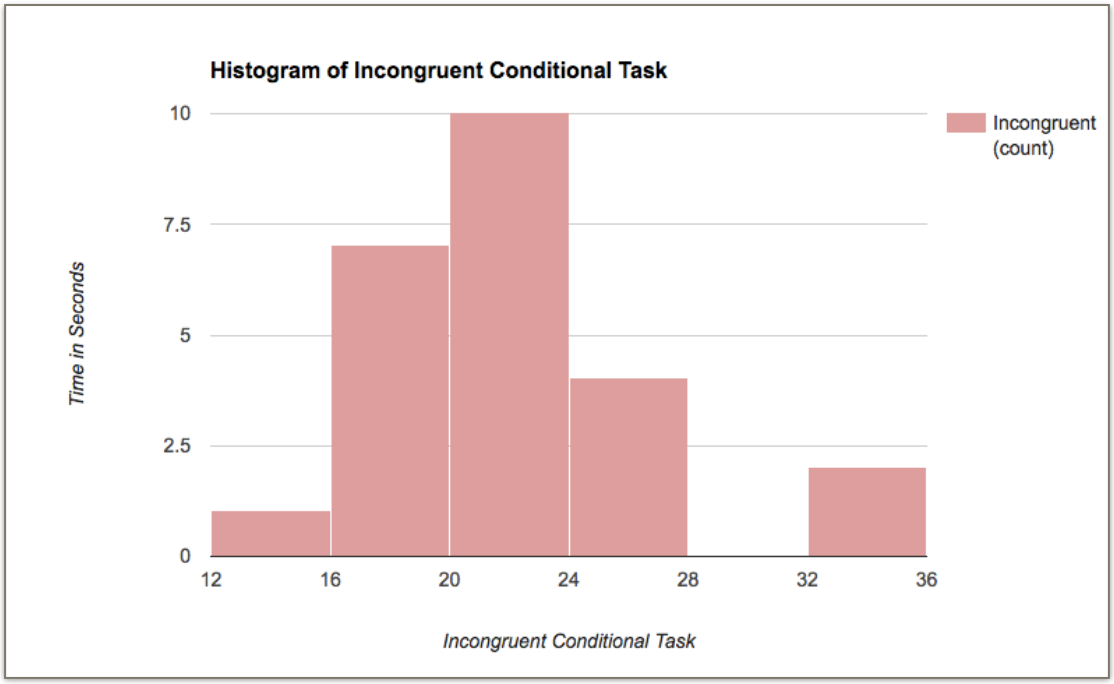


Figure 2: Histogram showing the time taken for Incongruent Conditional Task with bin size 4

In figure 3, I try visualize two the tasks using Box Plots. As it can be clearly seen, Incongruent task has a higher range and median than Congruent task (this gives an indication towards alternative hypothesis and rejection of null hypothesis).

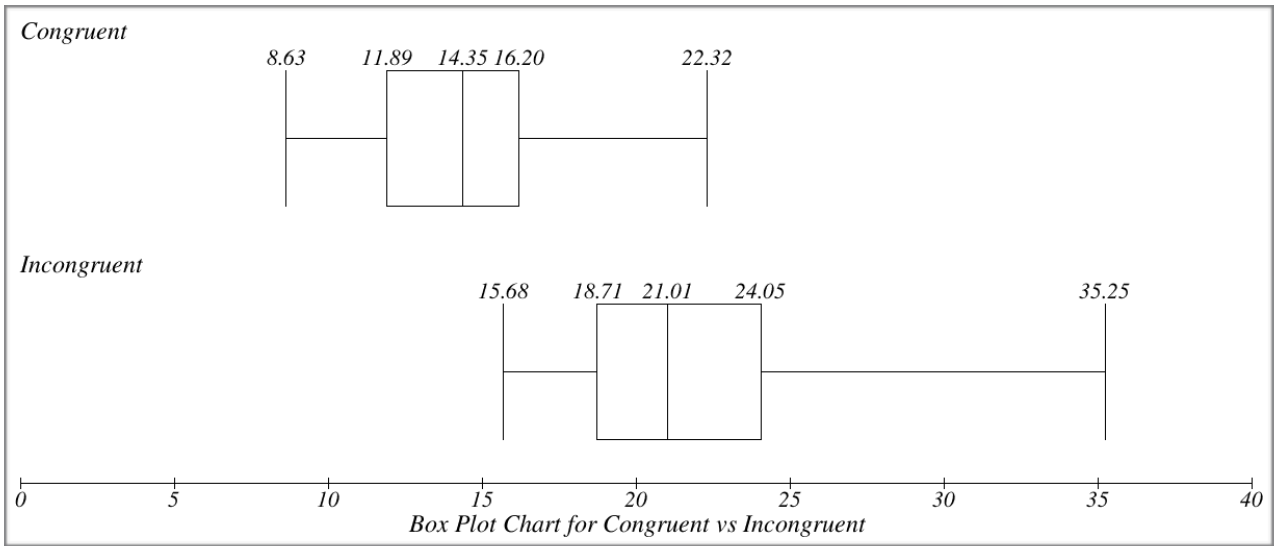


Figure 3: Box Plot Chart to compare the time taken for Congruent Vs Incongruent Conditional Task

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?

A. Following is the experimental parameters:

Type of test: Two-Tailed test
Alpha Level: .05

Calculation of $t_{critical}$

Degree of Freedom: (n-1) where n is size of sample.
Since size of sample is 24, the degree of freedom is 23.

Using the t-table, we select the value for .025 column and degree of freedom 23, $t_{critical}$ value will be (± 2.069).

Calculation of $t_{statistical}$

$t_{statistical} = (\mu_{congruent} - \mu_{incongruent})$ / Standard Error

where,

$\mu_{congruent}$ is the average of the time taken by each participant to name the ink colors in congruent word condition task; and
$\mu_{incongruent}$ is the average of the time taken by each participant to name the ink colors in congruent word condition task;

We take the difference of values for congruent and incongruent task and then calculate its standard deviation.

Standard Error = standard deviation for difference / √(sample size)

Mean time for congruent task $\mu_{congruent}$ = 14.05
Mean time for incongruent task $\mu_{incongruent}$ = 22.02
Standard Deviation for difference = 4.86

Standard Error = 4.86 / √24 = 0.993

$t_{statistical}$ = (14.05 - 22.02) / 0.993
$t_{statistical}$ = (-8.02)

Calculation of **Confidence Interval**

Alpha level .05 is equivalent to 95% confidence interval(CI):

Confidence Interval = ($\mu_{congruent}$ - $\mu_{incongruent}$) ± $t_{critical}$ * **Standard Error**
Confidence Interval = (-10.0225 , -5.9175)

Decision

With the value of $t_{statistical}$ is (-8.02) and Degree of Freedom being 23, **p value** is 0.001. This is less than alpha level .05. Hence we **reject the Null Hypothesis**.

The results show that the time taken for congruent conditional task is statistically significant than the time taken for incongruent conditional task.

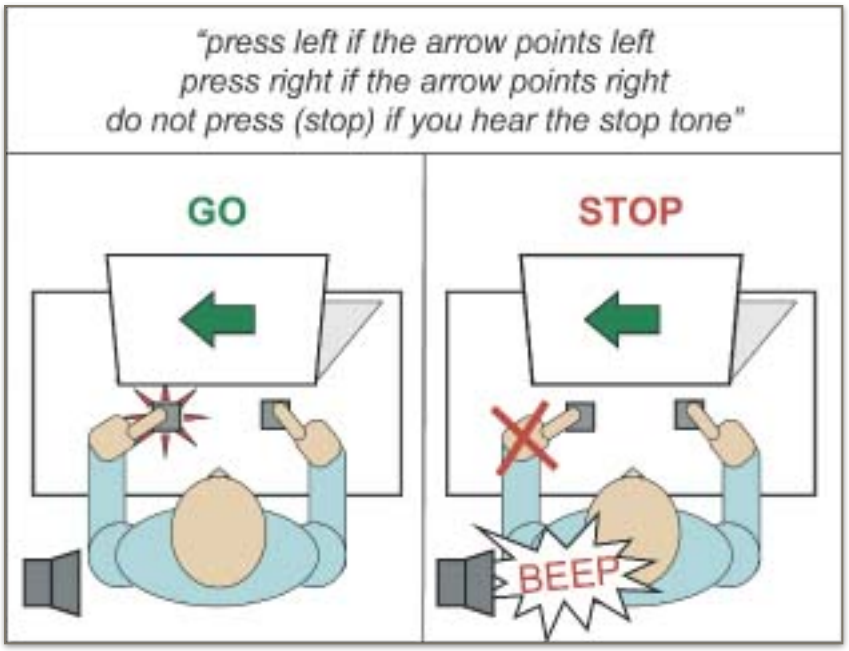
6. Optional: What do you think is responsible for the effects observed? Can you think of an alternative or similar task that would result in a similar effect? Some research about the problem will be helpful for thinking about these two questions!

A. The effects observed are due to stroop effect. After reading about the troop effect, a similar theory could be **Parallel distributed processing** theory.

This theory suggests that as the brain analyzes information, different and specific pathways are developed for different tasks. Some pathways, such as reading, are stronger than others, therefore, it is the strength of the pathway and not the speed of the pathway that is important. In addition, automaticity is a function of the strength of each pathway, hence, when two pathways are activated simultaneously in the Stroop effect, interference occurs between the stronger (word reading) path and the weaker (color

naming) path, more specifically when the pathway that leads to the response is the weaker pathway.
- Source: Wikipedia

An example could be how our brain works with the placement of buttons with respect to view while using a computers. If image direction and button placement is same(Congruent) then our brain respond faster, whereas if this is not the case(incongruent) then our brain takes time to process this.



References:
1. https://en.wikipedia.org/wiki/Stroop_effect
2. http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/null-and-alternative-hypotheses/
3. http://www.imathas.com/stattools/boxplot.html
4. https://en.wikipedia.org/wiki/One-_and_two-tailed_tests
5. http://www.socscistatistics.com/pvalues/tdistribution.aspx
6. http://dimensional-overlap.com/