Analyzing the Stroop Effect

We Perform the analysis in the space below according to the instructions (https://docs.google.com/document/d/1-OkpZLjG_kX9J6LIQ5IltsqMzVWjh36QpnP2RYpVdPU/pub?embedded=True) and the project rubric (https://review.udacity.com/#!/rubrics/71/view).

(1) What is the independent variable? What is the dependent variable?

Answer

The independent variable is the color of the ink in which the word is printed, whereas dependent variable is the time it takes to name the ink colors.

(2) What is an appropriate set of hypotheses for this task? Specify the null and alternative hypotheses, and clearly define any notation used. Justify the choices.

Answer

The hypothesis regarding current test is that incongruent words requires more time for recognition as the words displayed are color words whose names do not match the colors in which they are printed, which call for extra efforts to tell.

The null hypothesis is that the average time spent on identifing the colour of incongruent words will be no longer than that of congruent words, and the alternative hypothesis would be the average time spent on identifing the colour of incongruent words will be longer than that of congruent words.

H_0 : There is no difference in population means of response time under incongruent and congruent conditions ($H_0 : \mu C = \mu I$).

H_1 : Population mean of the response time under incongruent condition will be significantly larger than the response time under congruent condition ($H_1 : \mu C < \mu I$).

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/null-and-alternative-hypotheses/ (http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/null-and-alternative-hypotheses/)

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/what-is-a-hypothesis-test/ (http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/what-is-a-hypothesis-test/)

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/directional-and-nondirectional-hypotheses/ (http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/directional-and-nondirectional-hypotheses/)

(3) Report some descriptive statistics regarding this dataset. Include at least one measure of central tendency and at least one measure of variability. The name of the data file is 'stroopdata.csv'.

In [1]:

```
# Perform the analysis here
import pandas as pd
import numpy as np
from scipy import stats
import random
random.seed(42)
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
df = pd.read_csv('stroopdata.csv')
df.head()
```

Out[1]:

	Congruent	Incongruent
0	12.079	19.278
1	16.791	18.741
2	9.564	21.214
3	8.630	15.687
4	14.669	22.803

In [2]:

```
# Lower case of column name
df.columns = map(str.lower, df.columns)

# Find the mean
df.congruent.mean(), df.incongruent.mean()
```

Out[2]:

```
(14.051124999999999, 22.015916666666666)
```

In [3]:

```
# Find the median
df.congruent.median(), df.incongruent.median()
```

Out[3]:

```
(14.3565, 21.0175)
```

In [4]:

```
# Find the mode
df.congruent.mode()[0], df.incongruent.mode()[0]
```

Out[4]:

```
(8.6300000000000008, 15.687000000000001)
```

In [5]:

```
# Find the min and max

df.congruent.min(), df.incongruent.min(), df.congruent.max(), df.incongruent.max
()
```

Out[5]:

```
(8.6300000000000008,
 15.687000000000001,
 22.328000000000003,
 35.255000000000003)
```

In [6]:

```
# Find the variance
df.congruent.var(), df.incongruent.var()
```

Out[6]:

```
(12.669029070652176, 23.011757036231884)
```

In [12]:

```
# Find STD
np.std(df.congruent, ddof = 1), np.std(df.incongruent, ddof = 1)
```

Out[12]:

```
(3.5593579576451955, 4.7970571224691376)
```

Answer

Compare the central tendency of both columns, incongruent group possess relatively higher values, which are mean: 22.02, median: 21.01, mode: 15.69, variance: 23.01, STD: 4.79, as opposed to congruent group, mean: 14.05, median: 14.36, mode: 8.63, variance: 12.70, STD: 3.55.

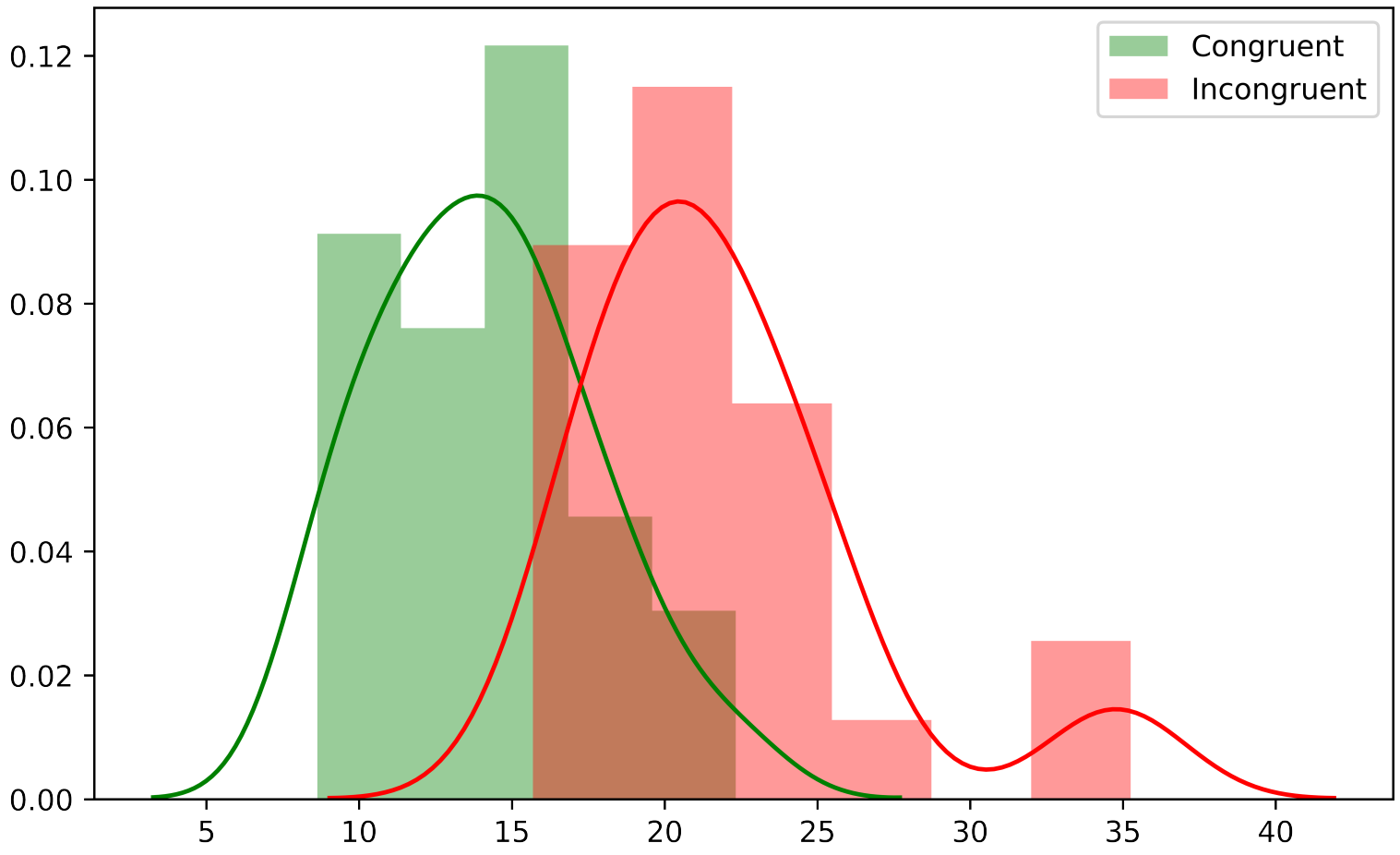
In addition to that, for min and max values, incongruent group present the numbers of 15.69 and 35.26, respectively, whereas congruent group suggests lower figures at 8.63 and 22.33.

(4) Provide one or two visualizations that show the distribution of the sample data. Write one or two sentences noting what we observe about the plot or plots.

In [8]:

```
# Build the visualizations here
fig, ax = plt.subplots(1, 1, figsize = (8, 5), dpi=1000)
# sns.distplot(df.congruent, color = 'g');
# sns.distplot(df.incongruent, color = 'r');

sns.distplot(df.congruent, color = 'g', label="Congruent");
sns.distplot(df.incongruent, color = 'r', label="Incongruent");
plt.legend();
ax.set_xlabel('');
```



Answer

The two groups fall into normal distribution, which is symmetric. The center of incongruent group is to the right of congruent group, and the variance of both groups looks similar.

(5) Now, perform the statistical test and report our results. What is our confidence level or Type I error associated with our test? What is our conclusion regarding the hypotheses we set up? Did the results match up with our expectations?

Regarding what test to use

T-score and Z-score are ideal candidates for currnt study.

T-score vs. Z-score: When to use a t score? The general rule of thumb for when to use a t score is when our sample:

> Has a sample size below 30;
>
> Has an unknown population standard deviation;
>
> We must know the standard deviation of the population and our sample size should be above 30 in order for us to be able to use the z-score. Otherwise, use the t-score.

So, I chose T_Score for current study.

> There are two types of T test, namely paried T test and paired T test. Paired T test is chosen for current study, the reason is "Paired samples" are when observations are made on pairs of units which are similar in some respect. Usually one treatment is applied to one member of each pair and not to the other which serves as the control. Pairing (or matching as it is sometimes called) can be done on the basis of age, sex, behaviour or any other factor that might be expected to have an effect on the response variable. The purpose of pairing is to reduce the variability in the response variable that we are measuring. The more similar the two individuals are, the more effective the pairing.

According to the plot above, the sample size seems on a normal distribution, and according to central limit therom, given random and independent samples of N observations each, the distribution of sample means approaches normality as the size of N increases, regardless of the shape of the population distribution. So as long as there is big enough sample, the distribution of population is supposed to be normal.

http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/t-score-vs-z-score/ (http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/t-score-vs-z-score/)

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/directional-and-nondirectional-hypotheses/ (http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/directional-and-nondirectional-hypotheses/)

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/what-is-a-critical-value/ (http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/what-is-a-critical-value/)

https://en.wikipedia.org/wiki/Student%27s_t-test#Unpaired_and_paired_two-sample_t-tests (https://en.wikipedia.org/wiki/Student%27s_t-test#Unpaired_and_paired_two-sample_t-tests)

http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%201/I.07%20normal.pdf (http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%201/I.07%20normal.pdf)

http://influentialpoints.com/Training/paired_t-test-principles-properties-assumptions.htm (http://influentialpoints.com/Training/paired_t-test-principles-properties-assumptions.htm)

In [11]:

```
# Perform the statistical test here

stats.ttest_rel(df.congruent, df.incongruent)
```

Out[11]:

```
Ttest_relResult(statistic=-8.020706944109957, pvalue=4.1030005857111
781e-08)
```

Answer

Type I error associated with current test is 0.05, I set this threshold mainly refered to most studies, and the conclusion is to reject the null hypothesis as the P value is far below the threshold. Such result matches my reasonable expection.

(6) Optional: What is responsible for the effects observed? Can we 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!

Answer

1. The difference between words and theirs background is responsible for the effects observed;

1. Similar task would be like, if speaking telephone while driving would increase the traffic accidents?

Conclusion

According to the statistical analyze, we reject the Null Hypothesis, which is to say the time spent on Incongruent condition is significantly higher than that of Congruent condition.

It makes sense, as from my own anecdotal experience incongruent words requires more time for recognition as the words displayed are color words whose names do not match the colors in which they are printed.