# StroopProject

*Shui(Sheena) Yu*

## Contents



### Background Information

In a Stroop task, participants are presented with a list of words, with each word displayed in a color of ink. The participant's task is to say out loud the color of the ink in which the word is printed. The task has two conditions: a congruent words condition, and an incongruent words condition. In the congruent words condition, the words being displayed are color words whose names match the colors in which they are printed: for example RED, BLUE. In the incongruent words condition, the words displayed are color words whose names do not match the colors in which they are printed: for example PURPLE, ORANGE. In each case, we measure the time it takes to name the ink colors in equally-sized lists. Each participant will go through and record a time from each condition.

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

**Independent Variable**: Condition of words congruence, one being a congruent words condition, the other being an incongruent words condition.

**Dependent Variable**: The time it takes the same participant to name the ink colors under each condition.

---

### 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 expect to perform a two-tail **paired samle t-test** in this case for the following two reasons:
  - First, based on the dataset we have, there are 24 participants in total, each of whom went through and recorded a time from *congruent words condition* and *incongruent words condition*. Different treatments (Stroop task under congruent words condition and Stroop task under incongruent

words condition) are applied to the same subjects (same 24 participants), thus we can say that these two sets of time data are dependent with each other, which means they are *paired data*. To analyze the paired data, it is often more reasonable to meansure the difference of each pair of record. Because our goal here is to test for any significant differences in the time recorded under two conditions by each participant.

  - Second, we are only given two groups of small sample data, without having a knowledge of the population parameters such as population variance. So we choose t-test to address the uncertainty of the standard error estimate.

- **Hypotheses:**

$$H_0 : \mu_{diff} = 0$$

$$H_A : \mu_{diff} \neq 0$$

The **Null Hypothesis** assumes that there is no difference between the mean time used under *congruent words condition* and *incongruent words condition*.

The **Alternative Hypothesis** assumes that there is difference between the mean time used under *congruent words condition* and *incongruent words condition*.

---

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

- **Read data and add the column of difference between Congruent values and Incongruent values**

```r
# Read data
setwd('/Users/Sheena/Udacity/Project1/')
stroopdata <- read.csv('stroopdata.csv', header = TRUE)

# Calculate the difference between Incongruent and Congruent values
stroopdata <- stroopdata %>%
    mutate(difference = Incongruent - Congruent)

knitr::kable(stroopdata)
```

| Congruent | Incongruent | difference |
|---:|---:|---:|
| 12.079 | 19.278 | 7.199 |
| 16.791 | 18.741 | 1.950 |
| 9.564 | 21.214 | 11.650 |
| 8.630 | 15.687 | 7.057 |
| 14.669 | 22.803 | 8.134 |
| 12.238 | 20.878 | 8.640 |
| 14.692 | 24.572 | 9.880 |
| 8.987 | 17.394 | 8.407 |
| 9.401 | 20.762 | 11.361 |
| 14.480 | 26.282 | 11.802 |
| 22.328 | 24.524 | 2.196 |
| 15.298 | 18.644 | 3.346 |
| 15.073 | 17.510 | 2.437 |

| Congruent | Incongruent | difference |
|---|---|---|
| 16.929 | 20.330 | 3.401 |
| 18.200 | 35.255 | 17.055 |
| 12.130 | 22.158 | 10.028 |
| 18.495 | 25.139 | 6.644 |
| 10.639 | 20.429 | 9.790 |
| 11.344 | 17.425 | 6.081 |
| 12.369 | 34.288 | 21.919 |
| 12.944 | 23.894 | 10.950 |
| 14.233 | 17.960 | 3.727 |
| 19.710 | 22.058 | 2.348 |
| 16.004 | 21.157 | 5.153 |

- **Measure of central tendency**

```r
# Calculate the mean and median of each group of data (Congruent, Incongruent, difference)
options(digits = 4)
measures_ct <- stroopdata %>%
    summarise_each(funs(mean, median), Congruent, Incongruent, difference)
```

Measures of central tendency inclue mean, median and mode of the specific data. We calculate means and medians of three groups of data in this case.

| Data Group | Mean | Median |
|---|---|---|
| Congruent | 14.0511 | 14.3565 |
| Incongruent | 22.0159 | 21.0175 |
| Difference | 7.9648 | 7.6665 |

- **Measure of variability**

Measures of variability include *range*, *interquartile range (IQR)*, *standard deviation (sd)* and so on. Here we will focus on these three metrics.

**Formula and explanation:** The *Range* is the simplest variability metric which measures the difference between the maximum data and the minimum data. The *IQR* stands for *Interquartile Range*, it measures the middle 50% of data in a distribution. Usually we get this range by substract first quartile from third quartile. The *Standard Deviation* is a measure of how spread out the data is in a distribution, it is also the square root of *Variance*. Notice here that in r, the function `sd` generates the sample standard deviation (denominator of variance formula is *n - 1*), which is suitable under this case because all we have is a sample, and we wish to make a statement about the population standard deviation from which the sample is drawn, we need to use the sample standard deviation.

$$Range = Max - Min$$

$$IQR = 75thPercentile(Q3) - 25thPercentile(Q1)$$

$$StandardDeviation = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

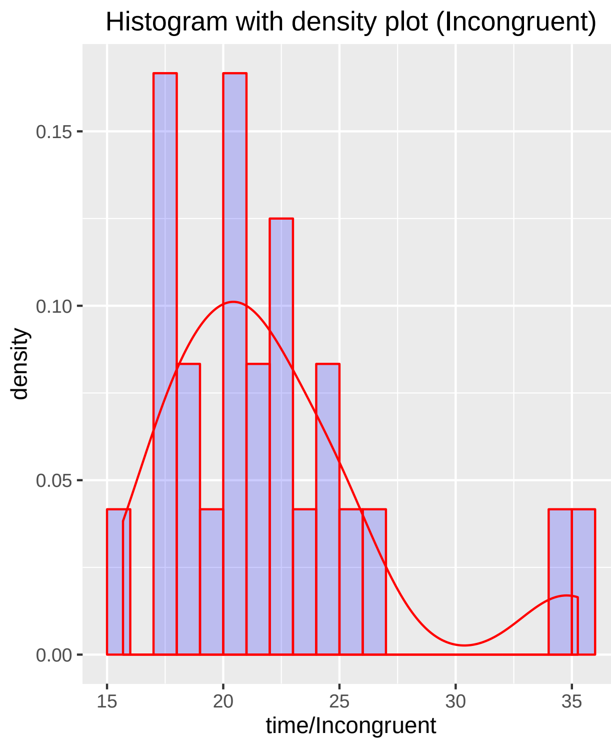
```r
# Calculate the range, IQR and standard deviation
range_var <- apply(stroopdata, 2, range)

IQR_var <- apply(stroopdata, 2, IQR)

sd_var <- apply(stroopdata, 2, sd)
```

| Data Group | Range | IQR | sd |
|---|---|---|---|
| Congruent | 13.698 | 4.3055 | 3.5594 |
| Incongruent | 19.568 | 5.3347 | 4.7971 |
| Difference | 19.969 | 6.613 | 4.8648 |

---

### 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.

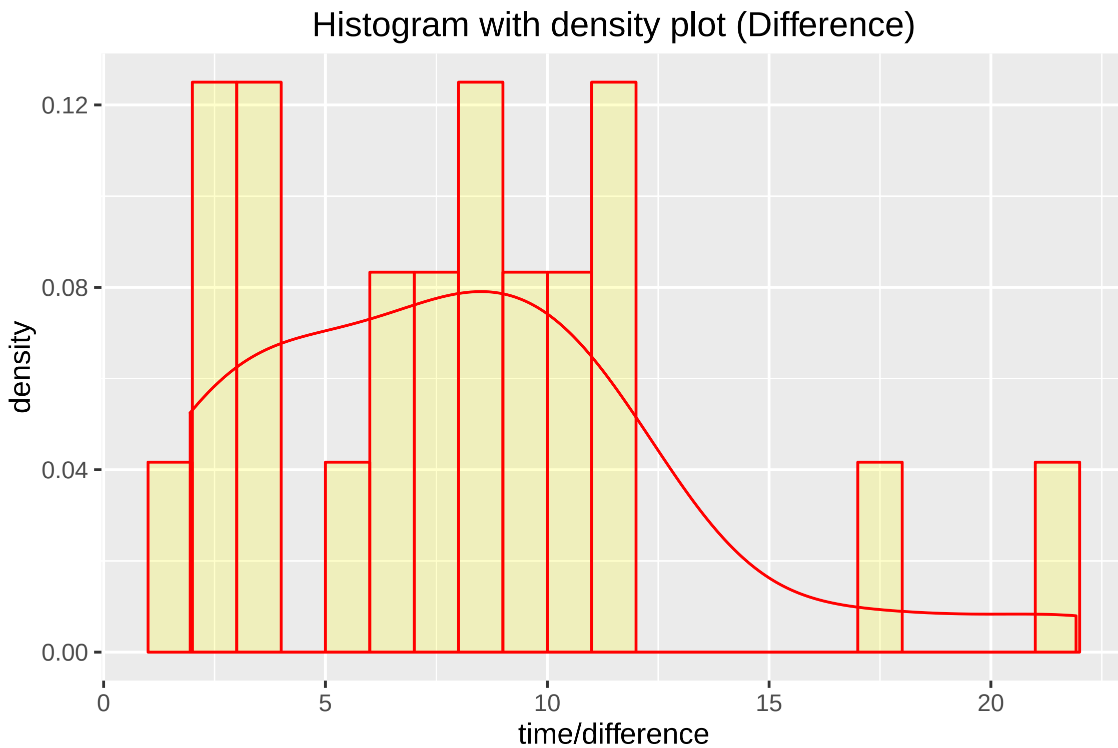
```r
# Distribution and histogram
dens_con <- ggplot(data = stroopdata, aes(stroopdata$Congruent)) +
    geom_histogram(aes(y = ..density..),
                   breaks = seq(8, 23, by = 1),
                   col = "red",
                   fill = "green",
                   alpha = .2) +
    geom_density(col = 2) +
    labs(title = "Histogram with density plot (Congruent)",
         x = "time/Congruent")

dens_incon <- ggplot(data = stroopdata, aes(stroopdata$Incongruent)) +
    geom_histogram(aes(y = ..density..),
                   breaks = seq(15, 36, by = 1),
                   col = "red",
                   fill = "blue",
                   alpha = .2) +
    geom_density(col = 2) +
    labs(title = "Histogram with density plot (Incongruent)",
         x = "time/Incongruent")
grid.arrange(dens_con, dens_incon, ncol = 2)
```


From the distribution plot of Congruent group of data, we can see that this is a unimodal distribution with most data values clustered around 14-15, and it has longer right tail.

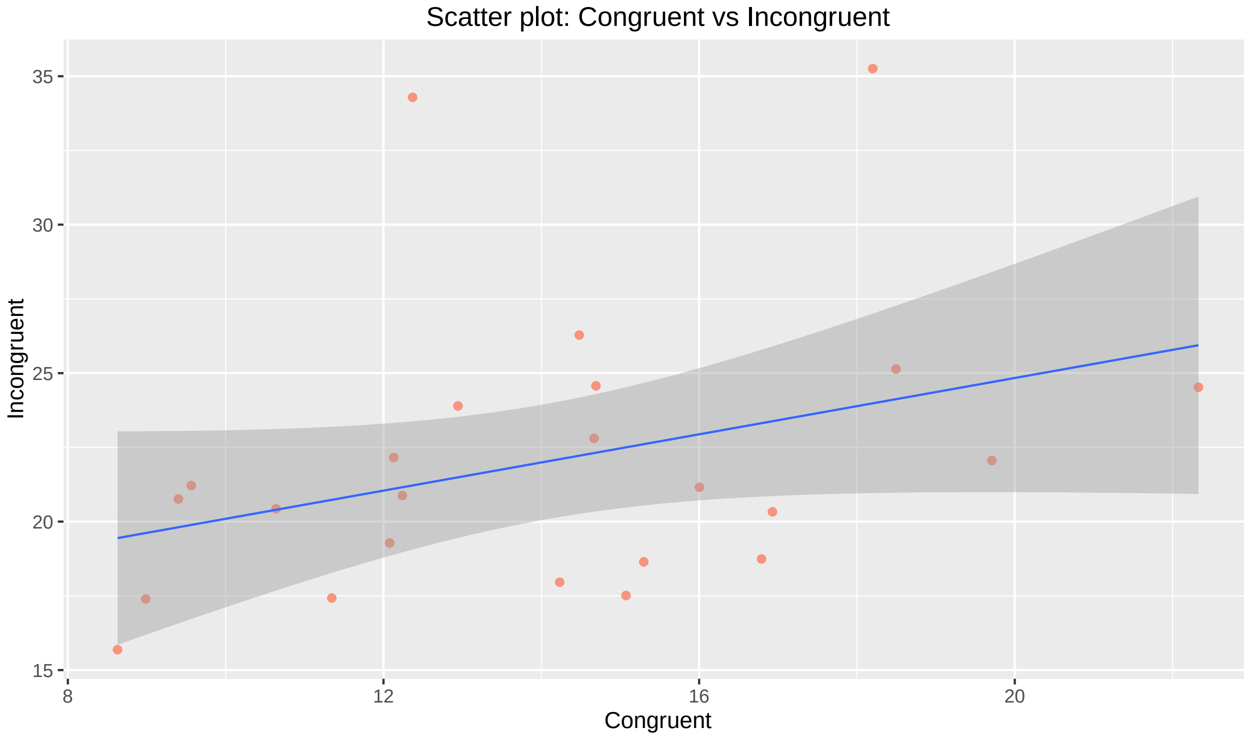
From the distribution plot of Incongruent group of data, we can easily see that this distribution has outliers on the far right-hand side, at around 35, so it looks a bit like bimodal shape but not quite. The majority of data values clustered around 18 to 22.

```
ggplot(data = stroopdata, aes(stroopdata$difference)) +
    geom_histogram(aes(y = ..density..),
                   breaks = seq(1, 22, by = 1),
                   col = "red",
                   fill = "yellow",
                   alpha = .2) +
    geom_density(col = 2) +
    labs(title = "Histogram with density plot (Difference)",
         x = "time/difference")
```




The most important data group in this case is the **difference** between `Incongruent` and `Congruent` data values, as we will be implementing **paired sample t-test** using this data group. From the chart above, we can see that there is no extreme peak in this distribution, with most data values lying on the left side of the plot and some outliers on the right. So this distribution has a long right tail.

```
# Scatter plot
ggplot(stroopdata, aes(x = Congruent, y = Incongruent)) +
    geom_point(alpha = 0.8, col = "#FD7D5F") +
    stat_smooth(method = "lm", formula = y ~ x, size = .5) +
    labs(title = "Scatter plot: Congruent vs Incongruent")
```




Let's take a deeper look at the **relationship** between time used under Congruent condition and time used under Incongruent condition for each participant using a *Scatter Plot*. The two dots on the upper side of the plot corresponds to the outliers in the *Histogram of Incongruent group of data with density plot*, indicating these two participants were not quite used to incongruent words condition and thus took longer time to identify the color of the ink in which the words are printed. Regardless of the outliers, the points follow a slightly upward trending line.

---

## 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?

- **Perform T-test**

```
# Calculate t critical value
T_lt <- qt(p = .025, df = 23)
T_rt <- qt(p = .975, df = 23)

# Calculate t-statistic
std_error <- sd_var[['difference']] / sqrt(24)
t_statistic <- (measures_ct[['difference_mean']] - 0) / std_error

# Calculate P-value
p_value <- pt(q = t_statistic, df = 23, lower.tail = FALSE) * 2
```

- **Confidence Level**

A confidence level refers to the percentage of all possible samples that can be expected to include the true population parameter. We set our confidence level at **95%** in this case.

- **Critical Statistic Value**

Our $H_0$ states that there is no difference between time used in congruent condition and that in incongruent condition at 5% significance level, while our $H_A$ assumes that there is difference between time used in these two conditions at 5% significance level.

What we do in this step is to find out the 2.5th ($t_{\alpha/2,n-1}$) and 97.5th ($t_{1-\alpha/2,n-1}$) percentiles of the Student t distribution with *23* (24 - 1) degrees of freedom.

The results are:

$$t_{(.025,23)} = -2.0687$$

$$t_{(.975,23)} = 2.0687$$

- **T-test results**

The *t-statistic* of the *difference* data group is 8.0207, which is much larger than the t critical value 2.0687 on the right-hand side tail. This finding also corresponds to our p-value 4.103e-08, which is pretty close to 0, meaning if in fact there is no difference between time used in congruent and incongruent condition, there is 4.103e-08 chance that a random sample of 24 participants would yield a sample mean difference of 7.96 or higher, which is very unlikely to happen simply by chance.

In conclusion, we reject the Null Hypothesis, and accept the Alternative Hypothesis that there is, indeed, difference between mean time used in both conditions for each participant at 0.05 significance level. The result match up with my expectations.

---

## 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!

Two theories developed by J. Ridley Stroop can account for this occurrence:

**1. Speed of Processing Theory**: He showed that the Interference occurs because, people are able to read words much quicker and to name the colour of the word is much more complex.

**2. Selective Attention Theory**: Interference occurs because; naming the actual colour of the words requires much more attention than reading the words.

A similar experiment would be "rotate words" task. It is implemented to test whether brain processing ability would be compromised by different shapes of the word.

## References


Sample Standard Deviation

OnlineStatBook

Confidence Level

Paired t-tests

The Stroop Effect

Different tests for the stroop effect