Proj1Statistics

January 4, 2017

```
In [1]: from IPython.display import display
```

1 "Statistics: The Science of Decisions"

This notebook is for submission to Udacity for the completion of the "Statistics: The Science of Decisions" - Project 1 of the Data Analyist Nanodegree.

Student Name: Arthur DeGraw

2 Background:

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.

3 Load pandas, math, etc libraries and the dataset

```
In [2]: import pandas as pd
        import math
        import matplotlib.pyplot as plt
        import matplotlib.style as style
        style.use('ggplot')
        %matplotlib inline
        from scipy.stats import ttest_rel, t
```

```
In [3]: stroop_df = pd.read_csv("stroopdata.csv")
        stroop_df.head()
```

```
Out[3]:    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
```

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

The independent variable are the classes of congruent/incongruent words and colors. The dependent variable is the average length of time to respond in each of the congruent/incongruent classes.

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

If mu_c and mu_i are the population average response times for congruent and incongruent classes respectively and mu_d=mu_i-mu_c then an appropriate set of hypotheses are:

null: H_0: mu_d = 0 alt: H_1: mu_d > 0

We will use an alpha level of 5% for this test.

That is: Does the incongruent class generally take longer to respond to than the congruent class? This choice was made as intuitively one would expect the test taker to have to filter the word and the color separately before responding in the incongruent class.

The test to be performed will be a t-test for difference of means with paired sample. This is because the same people took both tests, so the values are not independent between the two tests. Furthermore, the population standard deviation is not known so must be estimated from the sample data (t-test)

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

```
In [4]: stroop_df.describe()
```

```
Out[4]:         Congruent  Incongruent
        count  24.000000    24.000000
        mean   14.051125    22.015917
        std     3.559358     4.797057
        min     8.630000    15.687000
        25%    11.895250    18.716750
        50%    14.356500    21.017500
        75%    16.200750    24.051500
        max    22.328000    35.255000
```

The median time for response for the congruent and incongruent classes are 14.35 seconds and 21.02 seconds respectively. Seconds were assumed to be the units of measure since the webpage test reports time in seconds. The mean and standard deviation for the congruent class are 14.05

seconds and 3.56 seconds respectively. The longest time recorded in the sample for completing the tests are 22.33 seconds and 35.26 seconds respectively for the congruent and incongruent classes respectively.

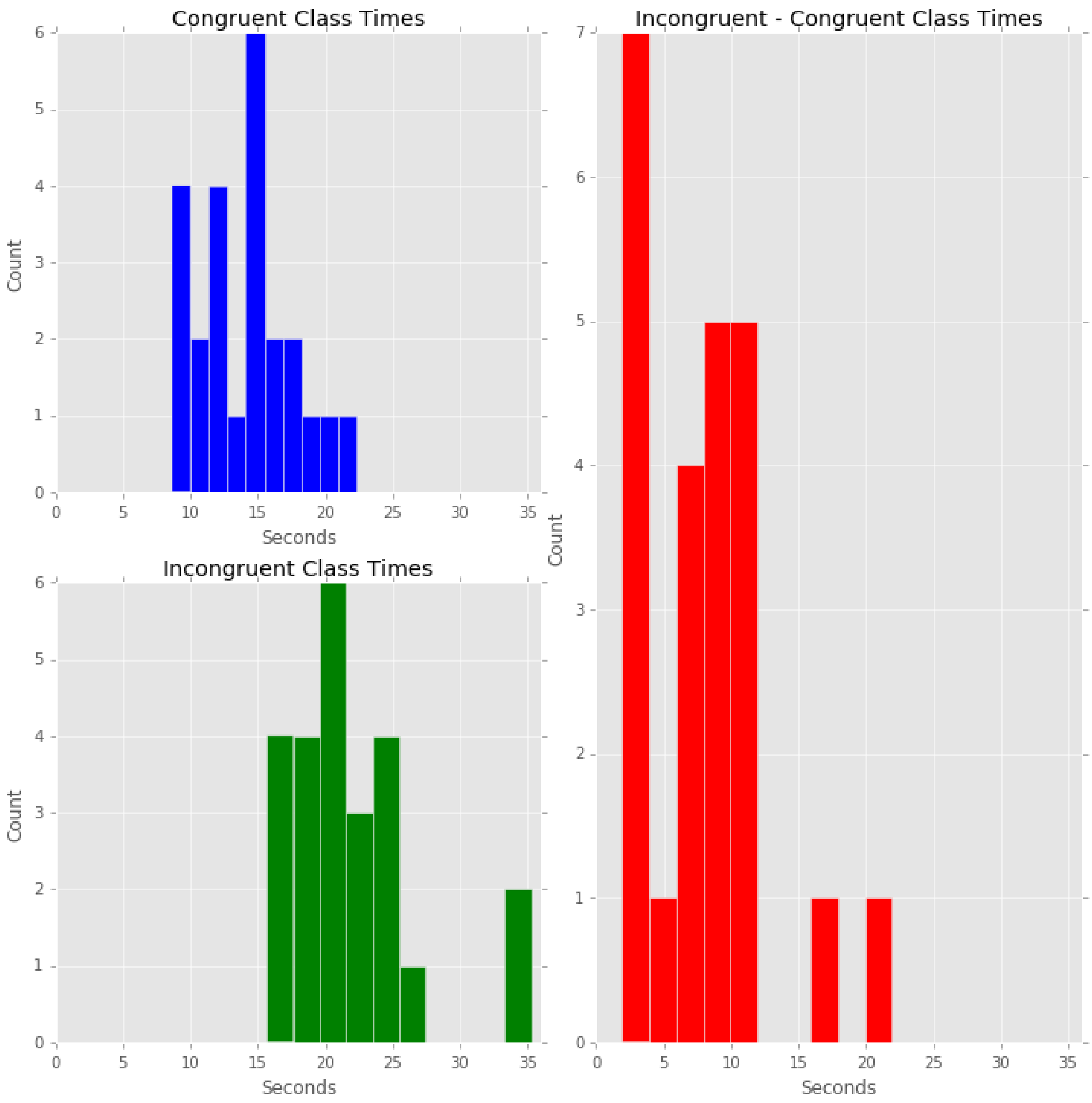
7 4. Provide one or two visualizations that show the distribution of the sample data. Write one or two sentences noting what you observe about the plot or plots.

8 Calculating the I-C values column

```
In [5]: stroop_df['diff(Inc-Con)'] = stroop_df['Incongruent'] - stroop_df['Congruer
```

9 Using matplotlib.pyplot to generate some histograms for visualization of distributions.

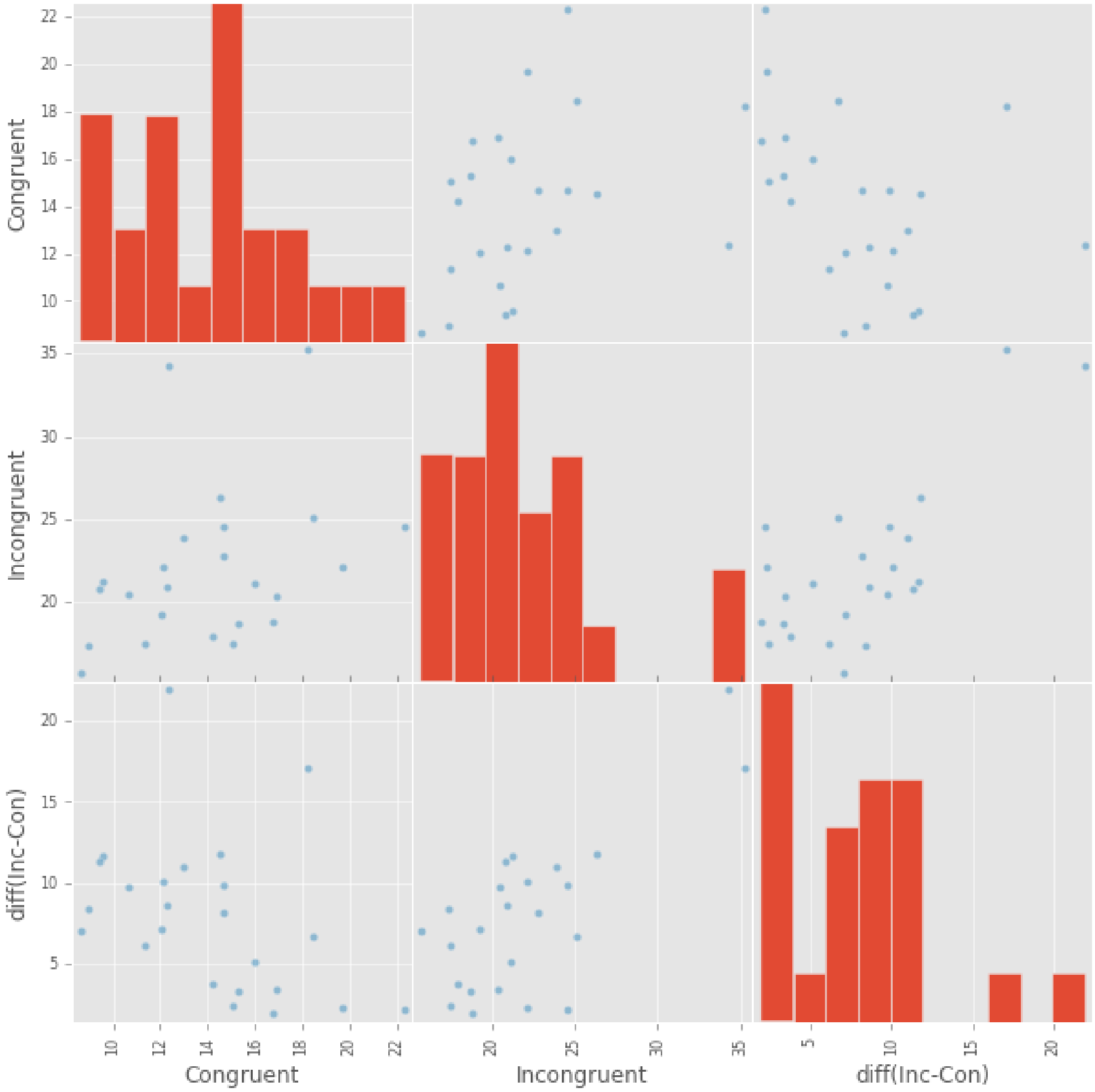
```
In [6]: fig=plt.figure(figsize=(10,10))
        ax=plt.subplot(221)
        ax.hist(stroop_df['Congruent'], color='blue')
        ax.set_xlim((0,36))
        ax.set_xlabel("Seconds")
        ax.set_ylabel("Count")
        ax.set_title("Congruent Class Times")
        ax1=plt.subplot(223)
        ax1.hist(stroop_df['Incongruent'], color='green')
        ax1.set_xlim((0,36))
        ax1.set_xlabel("Seconds")
        ax1.set_ylabel("Count")
        ax1.set_title("Incongruent Class Times")
        ax2=plt.subplot(122)
        ax2.hist(stroop_df['diff(Inc-Con)'], color='red')
        ax2.set_xlim((0,36))
        ax2.set_xlabel("Seconds")
        ax2.set_ylabel("Count")
        ax2.set_title("Incongruent - Congruent Class Times")
        plt.tight_layout(pad=0.4, w_pad=0.5, h_pad=1.0)
        plt.show()
```



The three histograms illustrate that the difference in times between the two tests are generally in the favor of Incongruent taking longer.

10 Exploratory data analysis with a scatter_matrix

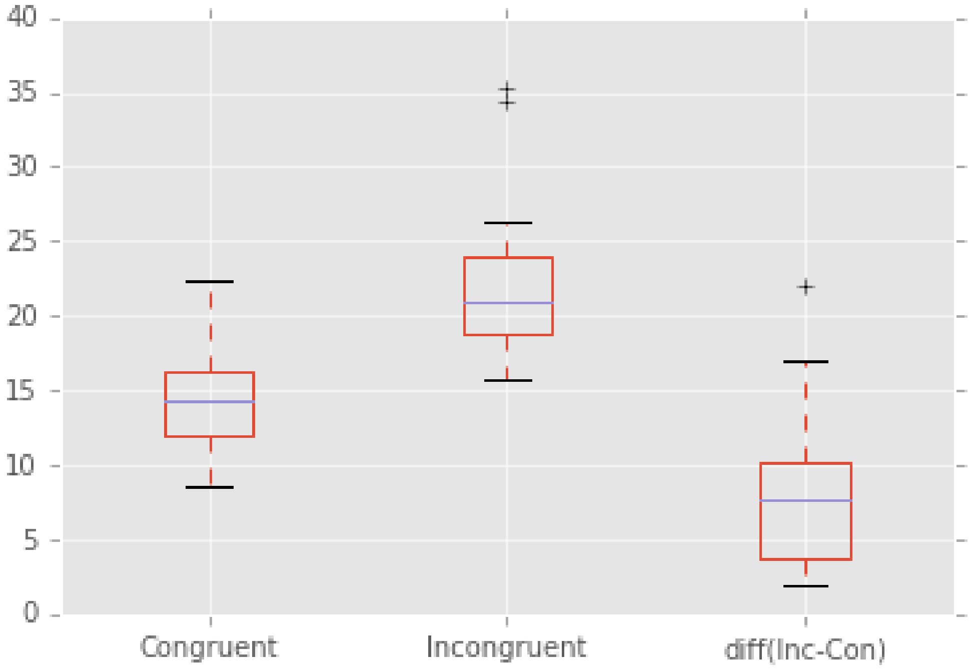
```
In [7]: plot = pd.scatter_matrix(stroop_df, figsize=(10,10), marker='o')
```



Not really anything to be gathered from the scatter plots.

11 And of course, we can look at a box plot to see how the sample data is spread as well. (Using dataframe method)

```
In [8]: box = stroop_df.boxplot(figsize=(5,5), return_type='dict')
```



The boxplots show a clear difference in the central 50% between the length of the tests for the Congruent and Incongruent classes.

12 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?

The test to be performed required computing the mean of the differences in times, the standard error for a paired t-test and the p-value.

```
In [9]: diff_mean = stroop_df['diff(Inc-Con)'].mean()
        print("The difference Incongruent-Congruent has mean %0.4f" % diff_mean)
        st_error = stroop_df['diff(Inc-Con)'].std() / (len(stroop_df) ** 0.5)
        print("The standard error for Incongruent-Congruent for the sample is %0.4f
        observed_t_score = (diff_mean - 0) / st_error
        print("The observed t value is %0.4f standard errors from 0" % observed_t_s

        #Since a right tail test then 1-cdf
        p = 1-t.cdf(observed_t_score, df=(len(stroop_df['diff(Inc-Con)']) - 1))
        print("The p-value for the hypothesis test is p= %.04e" % p)

        if p >= 0.05:
            print("Since p>alpha=0.05 we fail to reject the null hypothesis.")
        elif p<0.05:
            print("\nSince p<alpha=0.05 we reject the null hypothesis as there is s
        significant evidence to support the alternative hypothesis.")
```

```
The difference Incongruent-Congruent has mean 7.9648
The standard error for Incongruent-Congruent for the sample is 0.9930
The observed t value is 8.0207 standard errors from 0
The p-value for the hypothesis test is p= 2.0515e-08

Since p<alpha=0.05 we reject the null hypothesis as there is statistically signific
```

13 We could alternatively use the built in scipy t-test for related samples.

```
In [10]: t_val, p_val = ttest_rel(stroop_df['Incongruent'], stroop_df['Congruent'],
         print("Using Scipy's build in t-test for related samples: observed t = %0.
         p value = %.04e" % (t_val, p_val/2))
```

```
Using Scipy's build in t-test for related samples: observed t = 8.0207 ,p value = 2
```

```
In [ ]:
```

```
In [ ]:
```

14 References

http://pandas.pydata.org/pandas-docs/version/0.15.0/visualization.html
http://matplotlib.org/users/tight_layout_guide.html

```
In [ ]:
```