Student: Matthew Adam Mason
Date: April 30, 2017

Udacity Data Analyst
Project 1

1. Independent variable: Whether each word in a list of names of colors is printed in the same color as the word indicates — "word congruence".

   Dependent variable: The amount of time it takes to say aloud the colors in which words in a list are printed.

2. Let μ_c and μ_{ic} be the mean (population) amount of time to read aloud the colors in which words in a congruent list and an incongruent list, respectively, are printed. Then, we have the following hypotheses.

   The null hypothesis is that the population mean amount of time to read aloud the colors in which words in a congruent list are printed is equal to (or, more precisely, not significantly different from) the population mean amount of time to read aloud the colors in which words in an incongruent list are printed.

   The alternative hypothesis is that the population mean amount of time to read aloud the color of the word is significantly different between the congruent set and the incongruent set.

   H_0: $\mu_c = \mu_{ic}$
   H_A: $\mu_c \neq \mu_{ic}$

   We shall perform a t-test. While we may be able to calculate a mean and standard deviation for the population in our dataset, let's assume our dataset is a collection of samples from a larger population. Here, $n = 25$. This feels like a small sample, so a t-test seems appropriate.

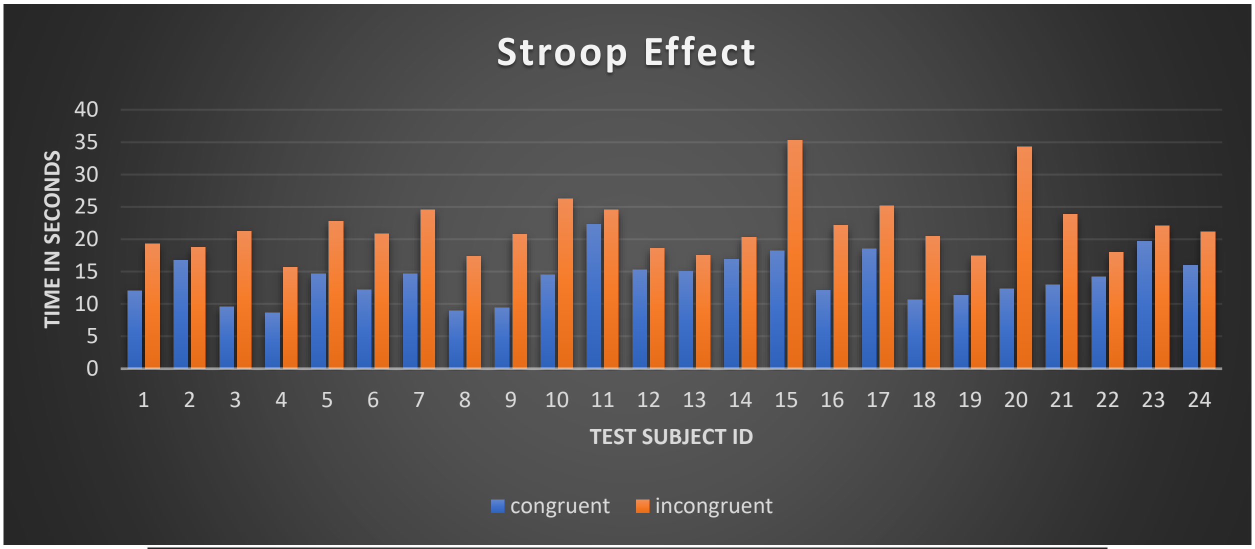
   This is an example of a dependent test. The same people were tested under the two conditions: congruent list and incongruent list.

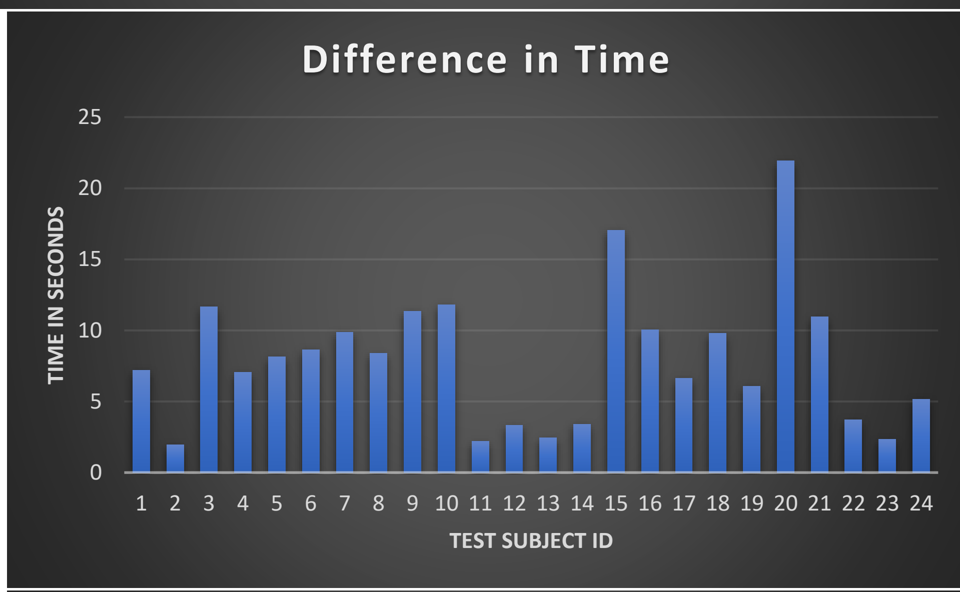
   Performing a t-test using dependent means involves some assumptions. First, the data is assumed to have an interval or ratio scale of measurement. Second, the data is assumed to be a random sampling from a defined population. Third, it is assumed samples or sets of data used to produce the difference scores are linked in the population through repeated measurement, natural association, or matching. Fourth, it is assumed scores, including difference scores, are normally distributed in the population.

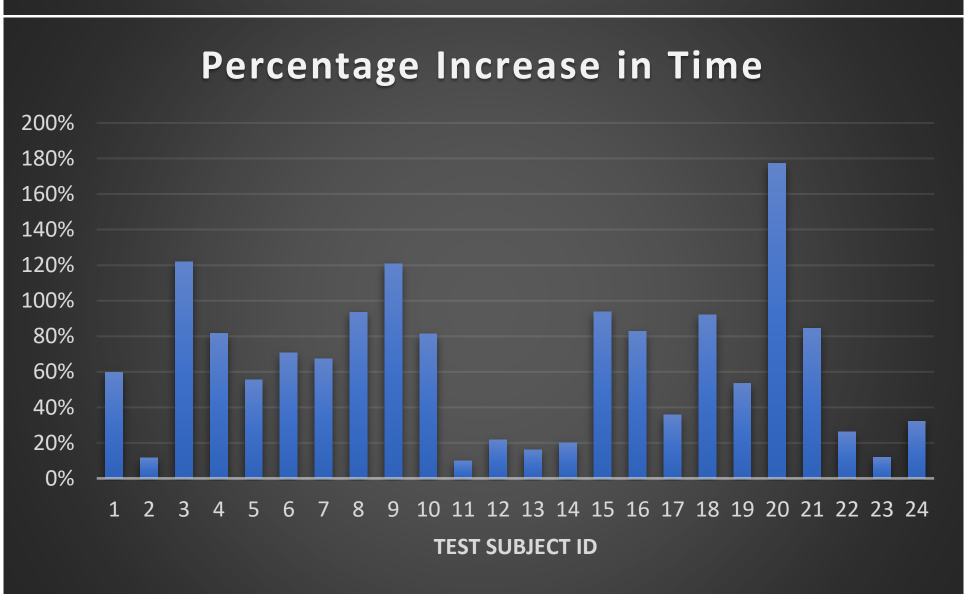
   Here, the data do not violate the assumptions. The scale of measurement is time in seconds, which has an absolute zero and equal intervals. Therefore, the data has a ratio (and hence interval) scale of measurement. Not much detail is provided regarding the characteristics of the test subjects sampled, so I assume the test subjects were randomly drawn from a general population of people who can read and speak in English. Third, the two sets—congruent and incongruent—are linked in the population in that each test subject has one score for the congruent set and another score for the incongruent set. Fourth, given the general population of people who can read and speak English and the nature of the test, we have no reason to believe the scores would be anything but normally distributed.

3. Let $\bar{x}_c$ and $\bar{x}_{ic}$ represent the congruent sample mean and the incongruent sample mean, respectively, and let $\bar{x}_d$ equal the mean difference in time between the congruent and incongruent tests.
   $\bar{x}_c = 14.051125$
   $s_c = 3.559357958$
   $\bar{x}_{ic} = 22.01591667$
   $s_{ic} = 4.797057122$
   $\bar{x}_d = 7.964791667$
   $s_d = 4.86482691$

4. We see from the charts that every test subject completed the test of congruent words faster. There was a noticeable difference in the range of percentage increase in time across test subjects.







5. $S_d = \frac{s_d}{\sqrt{n}} = \frac{4.86482691}{\sqrt{24}} = 0.993028635$

$t = \frac{\bar{x}_{ic} - \bar{x}_c}{S_d} = \frac{\bar{x}_d}{S_d} = \frac{7.964791667}{0.993028635} = 8.020706944$

We can see already that we reject the null hypothesis, regardless of α. Nevertheless, let $\alpha = .01$. Then, in a two-tailed test based on 23 degrees of freedom, $t_{\text{crit}} = 2.807$. We reject the null hypothesis because the t-statistic is in the critical region.

Let's calculate the 99% confidence interval.

$$CI = \bar{x}_d \pm t_{\text{crit}} \times S_d = 7.964791667 \pm 2.807 \times 0.993028635 = (5.177360289, 10.75222304)$$

In conclusion, the amount of time required to read aloud the colors in the incongruent set is greater than the amount of time required to read aloud the colors in the congruent set. With 99% confidence, the mean amount of time required by the underlying population to read aloud the colors in the incongruent set is between 5.177360289 and 10.75222304 seconds greater than the mean amount of time required by the underlying population to read aloud the colors in the congruent set. This met with my expectation given my performance on the test.

References:

Wikipedia articles on standard error and confidence interval.
"Dependent t-test for paired samples" available at https://statistics.laerd.com/statistical-guides/dependent-t-test-statistical-guide.php
"Independent t-test using SPSS Statistics" available at https://statistics.laerd.com/spss-tutorials/independent-t-test-using-spss-statistics.php
"Test Assumptions" available at http://www.psychology.emory.edu/clinical/bliwise/Tutorials/TOM/meanstests/assump.htm
"Levels of measurement" available at http://psc.dss.ucdavis.edu/sommerb/sommerdemo/scaling/levels.htm