Student:
Christophe Sebastien
Christophe.sebastien@airbus.com

Date:
Tuesday 29th of November 2016

Nanodegree:
Data Analyst Nanodegree

Project:
P1: statistics: the science of decisions [1]

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

Based on the project description [1] and the definition of dependent and independent data test [2]:

the Independent variable is the 'color of the ink in which the word is printed'.
the Dependent variable is the 'time it takes to name the ink colors in equally-sized lists'

2.1. What is an appropriate set of hypotheses for this task?

According to [3], *"the Stroop effect is the finding that naming the colors of color words (e.g. the words 'green', 'red', 'blue', etc.) is easier and quicker if the actual observed colors of the words match the colors that the words denote (e.g. the colors green, red, blue, etc., respectively) than if they do not match".*

Basically, the can be formulated as follow: does it take more time to identify incongruent data than congruent data set.

Then,
H_0: Null hypothesis:
It takes less time or there is no difference in the time it takes to name the ink colors between the congruent words condition list and the incongruent words condition list.

Ha: Alternative hypothesis:
It takes more time to name the ink colors between the congruent words condition list and the incongruent words condition list.

Let us define:

μ_C :the population mean of the time taken to identify the congruent data
μ_I : the population mean of the time taken to identify the incongruent data

Then:

H_0: $\mu_I - \mu_C \leq 0$ (1)
H_A: $\mu_I - \mu_C > 0$ (2)

2.2. What kind of statistical test do you expect to perform?

In the provided data [4], we do not know the entire population, only a sample, we need then to use statistical Student's t-test.

In this test, the same person is given two conditions to see how they react to each one. The provided samples are then dependent.

We want to know if there is increase in time, we will use then a one- tailed t-test.

3. Descriptive statistics including central tendency and variability.

In the provided data [4], we can compute:

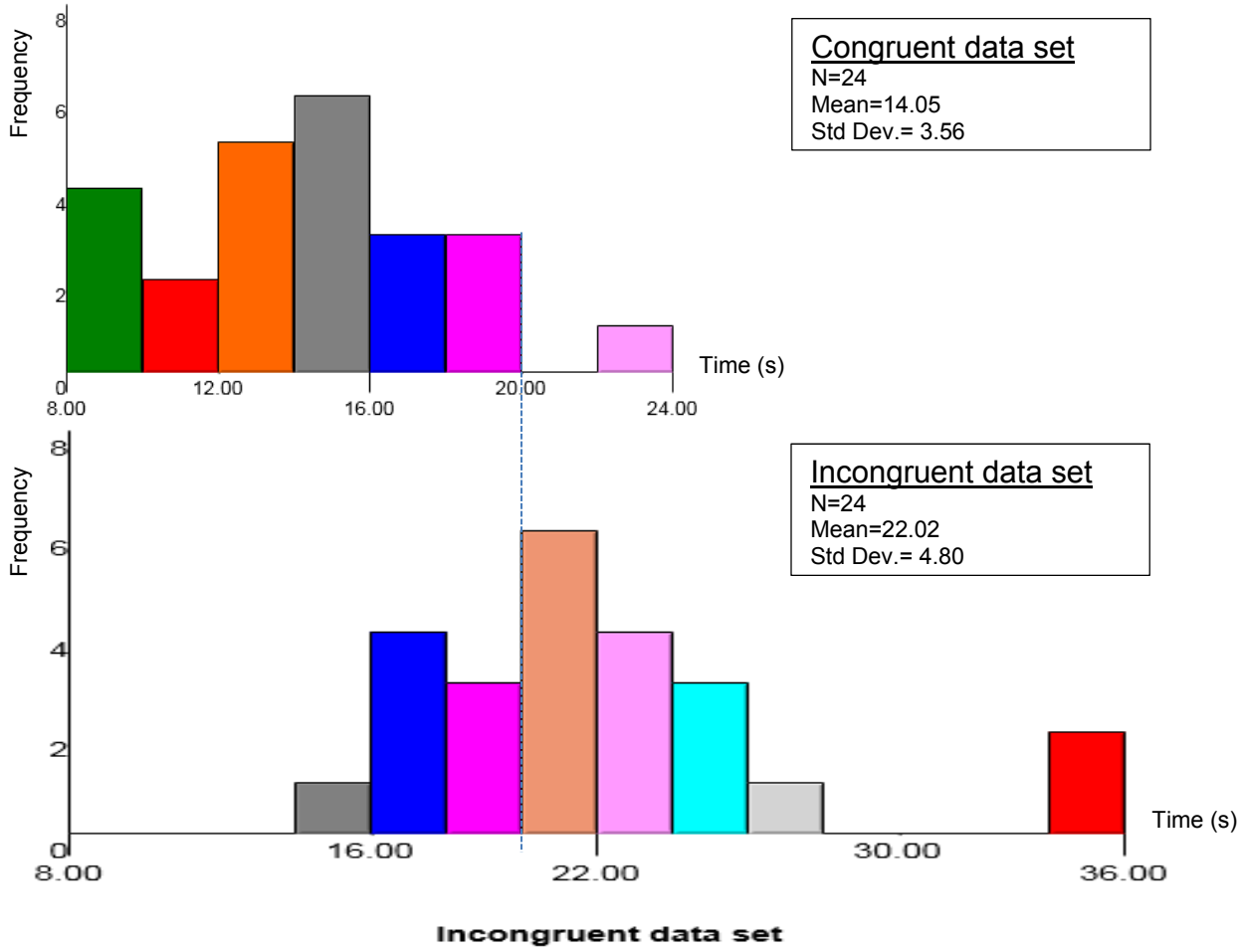
Congruent data set (_c)	Value	Note
n_c=	24	sample size
Mean_c ($\overline{X_c}$)=	14.05	sample mean
Median_c=	14.36	Median value of the sample
Range_c=	13.70	Max-Min of the sample
Variance (for a sample)_c (S_c^2)=	12.67	Variance for a sample: $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$
Standard Deviation (for a sample)_c (S_c)=	3.56	Standard deviation for a sample: $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}}$
Standard Error_c (SE_C)=	0.73	Standard deviation divided by the squared root of the sample size: $SE = {S}/{\sqrt{n}}$

Incongruent data set (_i)	Value	Note
n_i=	24	sample size
Mean_i ($\overline{X_i}$)=	22.02	sample mean
Median_i=	21.02	Median value of the sample
Range_i=	19.57	Max-Min of the sample
Variance (for a sample)_i (S_i^2)=	23.01	Variance for a sample: $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$
Standard Deviation (for a sample)_i (S_i)=	4.80	Standard deviation for a sample: $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}}$
Standard Error_i (SE_i)=	0.98	Standard deviation divided by the squared root of the sample size: $SE = {S}/{\sqrt{n}}$

Delta Incongruent-Congruent data	Value	Note
Point estimate M_D	7.96	Mean difference $M_D = \overline{X_i} - \overline{X_c}$
Standard Deviation (for a sample)_Δ (S_Δ)=	4.86	Standard deviation of the difference between sample means: $SE = {S}/{\sqrt{n}}$

4. Visualizations to show the distribution of the sample data. Interpretation:

Histogram frequency versus time



Note: histogram performed using [5]

Conclusion:

- For the congruent data and the incongruent data, the distributions look approximatively normal (even if more data would help to judge).
- For the Incongruent data have a higher mean value and the values are more dispersed (higher standard deviation)

5. Perform the statistical test. 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?

Statistical test:
As mentioned in chapter 2.2, we will perform a one- tailed t-test considering the hypothesis formulated in chapter 2.1.

From chapter 3, we have:

$M_D=\overline{X_i}-\overline{X_c}$=7.96: Point estimate or mean difference
S_Δ=4.86: Standard deviation of the difference between sample means
n=24: sample size

The t statistic is:

$$t_{statistic}=\frac{\overline{X_i}-\overline{X_c}}{S_\Delta/\sqrt{n}}$$

$\boldsymbol{t_{statistic}}$ **=8.02**

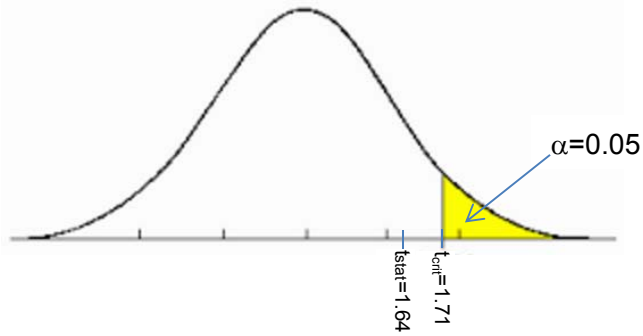
The t critical value is:

df=n-1
df=23 degrees of freedom
We will consider a α level of 0.05
According to the t statistic table,

$t_{critical}$(23)~1.714, p>0.05, one-tailed

Conclusion on H_0:

$t_{statistic}$ is superior to $t_{critical}$ at an α level of 0.05
We reject H_0



According to the sample data used in this test and the definition of the effect [3], we conclude that **the Stroop effect is verified**. It takes longer to identify incongruent data than congruent data.

Note: for a smaller a level of 0.001, $t_{critical}$(23)~3.485, p>0.001, one-tailed. The Stroop effect is even verified for a probability of 99.9%.

The confidence interval is:

$$CI = M_D \pm t_{critical}\left(S_\Delta / \sqrt{n}\right)$$

$\mathbf{95\%}\boldsymbol{CI} = [\mathbf{6.26}; \mathbf{9.67}]$ on the mean difference
We can conclude that on average people will take around 6 to 10seconds more time with incongruent data than congruent ones at a confidence of 95%.

Conclusion:
Subjectively, and after performing myself the test, I was expecting that it takes more time to properly identify incongruent data than congruent ones. The t-test is confirming this impression. The additional time of 6 to 10seconds to identify incongruent data versus congruent ones looks quite large.

6.1. Optional: What do you think is responsible for the effects observed?
According to [6],

> *"Two theories that may explain the Stroop effect:*
> *Speed of Processing Theory: the interference occurs because words are read faster than colors are named.*
> *Selective Attention Theory: the interference occurs because naming colors requires more attention than reading words."*

Subjectively, both of these theories make sense.

6.2. Can you think of an alternative or similar task that would result in a similar effect?
[6] propose a number of similar tests based on:

- picture of an animal with another animal name written on the picture
- position of a word in a square which does not necessary match the word meaning (top, left,…)

From my point of view, this effect shows, for example, the importance of well design car of aircraft cockpit. Imagine that an aircraft pilot have to think twice before moving the stick or pressing a button due to the badly organized cockpit or incoherent stick command?

References
[1]: Project description
https://docs.google.com/document/d/1-OkpZLjG_kX9J6LIQ5IltsqMzVWjh36QpnP2RYpVdPU/pub?embedded=True

[2]: Dependent and Independent variable
https://en.wikipedia.org/wiki/Dependent_and_independent_variables

[3]: Description of Stroop effect
https://en.wikipedia.org/wiki/Stroop_effect

[4]: Project data set
https://drive.google.com/file/d/0B9Yf01UaIbUgQXpYb2NhZ29yX1U/view

[5]: Histogram tool:
http://www.shodor.org/interactivate/activities/Histogram/

[6]: Information about Stroop effect
https://faculty.washington.edu/chudler/words.html