TEST A PERCEPTUAL PHENOMENON

GRADING PAPER FOR THE COURSE STATISTICS: THE SCIENCE OF DECISIONS, PART OF UDACITY'S NANODEGREE

Author: Nikita Barsukov, nikita@barsukov.net

INTRODUCTION

This is a write-up for the first course in Udacity's data analyst nanodegree. Here I was required to calculate basic descriptive statistics and perform hypothesis testing on a small dataset representing result from Stroop experiment, a psychological test.

QUESTIONS FOR INVESTIGATION

WHAT IS DEPENDENT AND INDEPENDENT VARIABLE?

There are two independent variable in the dataset: "Congruent" and "Incongruent", representing times to complete tasks for words that match the font color, and words that do not match color. There are no dependent variables in the dataset.

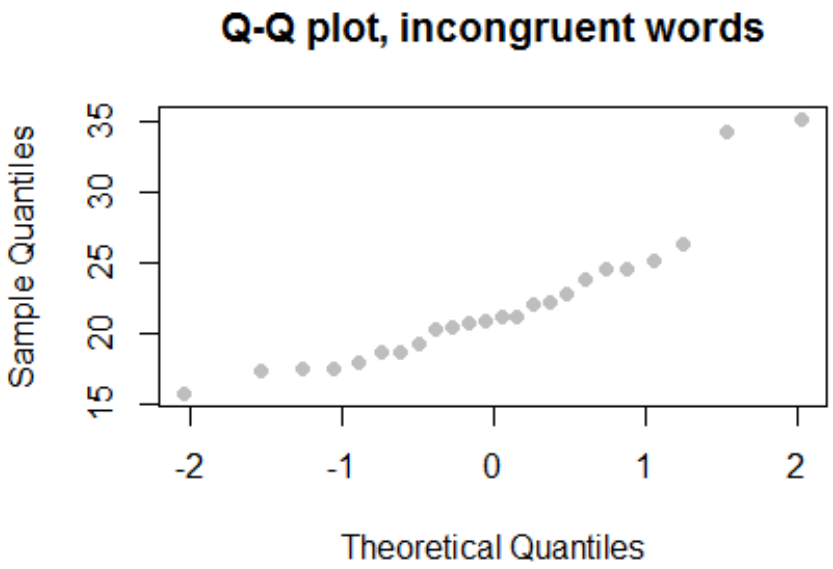
WHAT ARE APPROPRIATE SET OF HYPOTHESIS?

Dataset contains time that it took to name all the colors of words for a situation when colors were the same as words ('Congruent' variable), and when color was different that words ("Incongruent" variable). Our **null hypothesis** can be "Mean time to name all the colors of word was the same when words matched word color, and when words were not the same as their color". Alternative hypothesis would be "Mean time to name all the colors of words is different when words matched word color, and when words were not the same as their color".

Since the dataset is rather small, 24 observations in each dataset, we should be cautious in strictly determining if the distribution of our variables is random or not.

Creating histogram obviously would be meaningless with such a small sample, however let's construct Q-Q plot to see if our data is close to normal or not.

Two plots on Figure 1 below show us that on both plots dost are rather close to the straight line. Thus we could accept that the distribution of both our variables is close to normal. This means that we can use Student's t-test in our hypothesis testing.



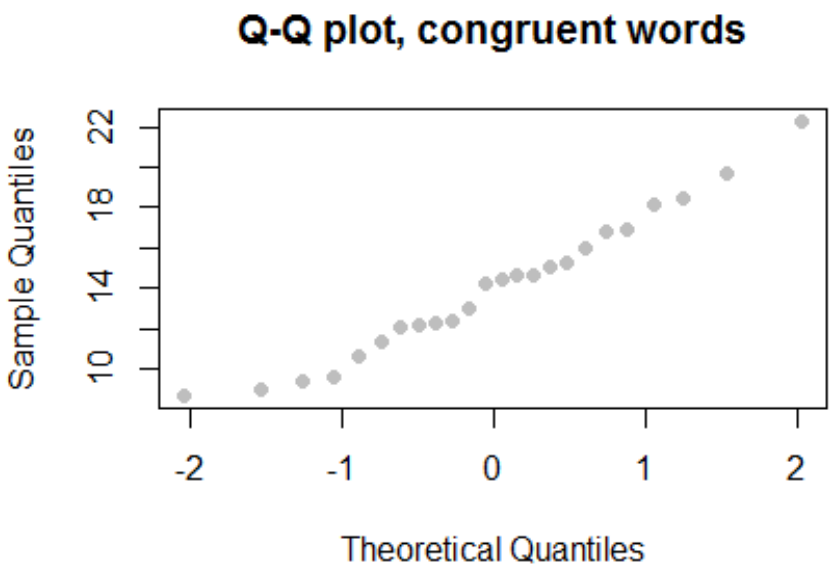


FIGURE 1

DESCRIPTIVE STATISTICS

Mean and standard deviation of two variables are displayed below:

Variable	*Mean*	*Standard Deviation*
Congruent	14.05113	3.559358
Incongruent	22.01592	4.797057

DATA VISUALIZATIONS

I used R programming language to generate boxplot and scatterplot of two variables. As we can see from boxplots in Figure 2 below, distributions differ visibly between each other. On the other hand number of observation is rather small, 24 for both variables. This can lead to inconclusive results from out statistical tests which I aim to perform in next chapter.

Scatterplot on Figure 3 below strongly suggest that these to variables are indeed independent, and we can indeed use T-test for our hypothesis tests.

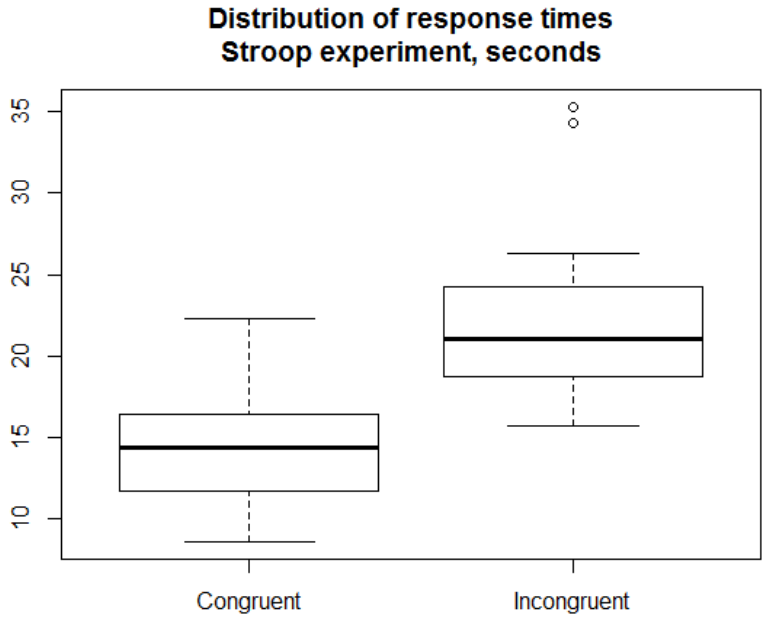


FIGURE 2

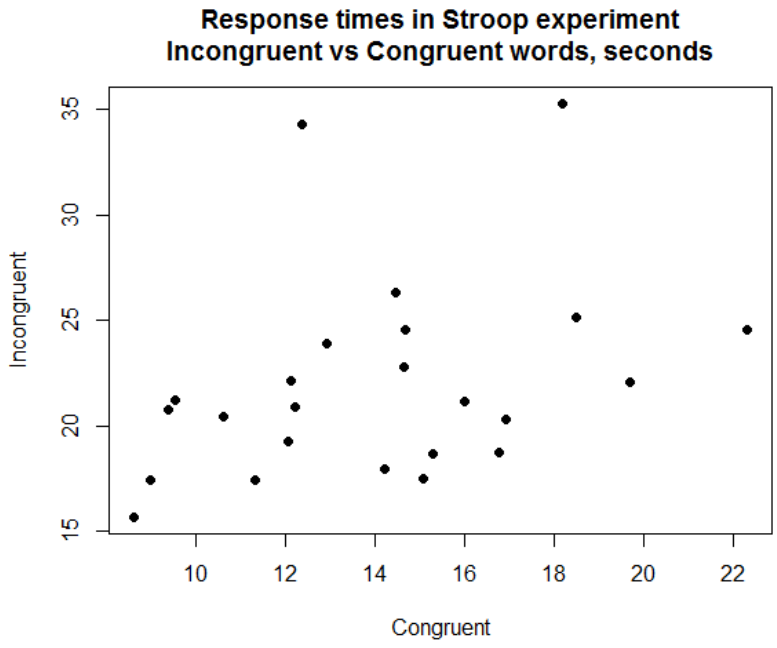


FIGURE 3

HYPOTHESIS TESTING

Two-sided t-test with confidence level 99% gave us following results:

Parameter	Value
t-score	-6.5323
critical t-score at 99%	-2.696762
p-value	~0%
99% confidence interval	-11.252959 – -4.676624

We can clearly see that with these results we can **reject null hypothesis** with great confidence. This means that within described experiment it takes longer time to correctly name all the colors of words, when words do not match their colors. This corroborates with the original description of this psychological experiment.

SOURCES

1. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
2. https://en.wikipedia.org/wiki/Normality_test
3. https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot
4. http://stackoverflow.com/questions/11526041/critical-t-values-in-r
5. http://www.statmethods.net/advgraphs/layout.html