Statistics: The Science of Decisions

Udacity Data Analyst Nanodegree – Project 1

Questions

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

The **independent variable is the condition of the words** which are presented to the subjects, which may either be congruent, e.g. red is written in red, or incongruent, e.g. red is written in green. The **dependent variable is the time it takes subjects to name the ink colors**. Alternatively, our dependent variable could be the amount of errors participants make when naming the ink colors.

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

The null hypotheses should be that there is no difference in mean response times when subjects are presented with congruent or incongruent conditions of words, thus: $\mathbf{H_0: \mu_c = \mu_i}$ – where:

- $\mathbf{H_0}$ denotes the null hypothesis
- $\mathbf{\mu_c}$ is the mean response time of participants who are dealing with congruent words
- $\mathbf{\mu_i}$ is the mean response time of participants who are dealing with incongruent words

Being conservative we may formulate the alternative hypothesis in two directions, i.e. the mean response time of subjects presented with congruent words can be higher **or** lower than the response time of subjects presented with incongruent words (or vice versa). Thus the alternative hypothesis should be: Subjects will show differences in mean response time when identifying congruent or incongruent labeled words: $\mathbf{H_A: \mu_c \neq \mu_i}$ – where:

- $\mathbf{H_A}$ denotes the alternative hypothesis
- $\mathbf{\mu_c}$ is the mean response time of participants who are dealing with congruent words
- $\mathbf{\mu_i}$ is the mean response time of participants who are dealing with incongruent words

Alternatively, we might intuitively hypothesize that we expect response time to be higher (or equal) if subjects have to identify incongruently labeled words in relation to congruently labeled words, thus: $H_A: \mu_c \leq \mu_i$ - **the following analyses will however use the two-directional alternative hypothesis**.

Since this is an experiment and we do not know the population parameters, we should conduct t-tests. More specifically, since the data for the project contains two results for each participant, i.e. the same participant took the test twice given two different conditions, we should conduct a **paired-sample t-test for dependent samples**. In order to conduct the test, the following assumptions have to be met:

1. Data must be interval or ratio scale
2. Sample must be done in a random way from a defined population

3. Samples to produce the difference scores are linked in the population through repeated measurement, natural association or matching
4. Scores are normally distributed

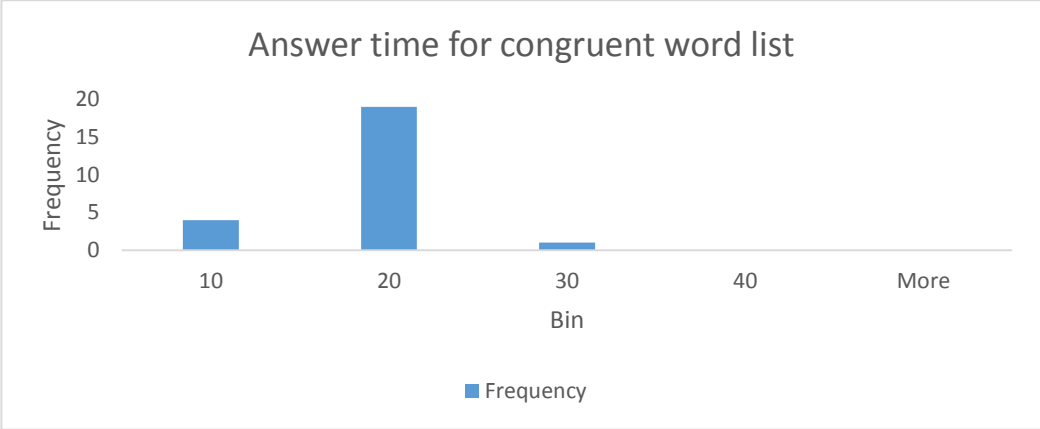
Looking at our sample we can immediately confirm that assumptions 1 (ratio data) and 3 (repeated measurement) are met. Since we have no insight into the sampling process, we assume that assumption 2 is met as well. Looking at the visualizations in section 4 we can further confirm that the data is approximately normally distributed and thus also confirm assumption 4. Even if the data were not following a normal distribution, the t-test for paired samples is robust enough for such violations.

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

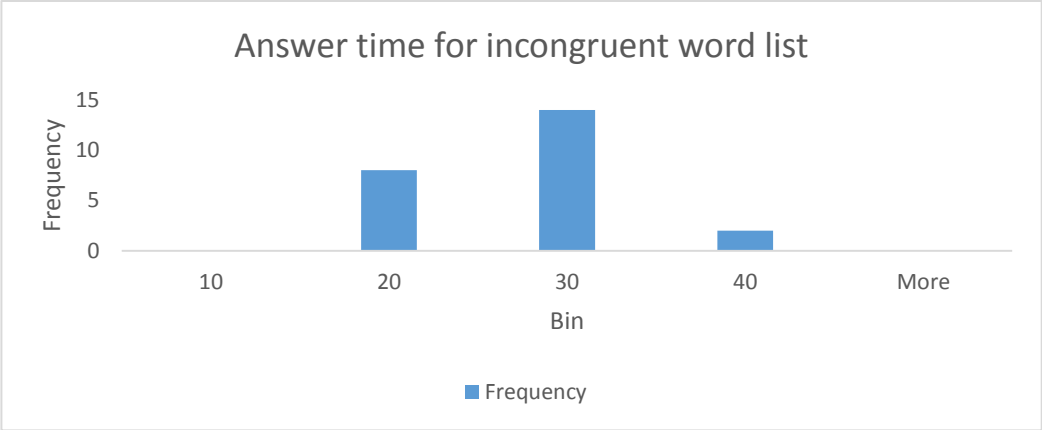
Descriptive statistics	**Congruent**	**Incongruent**
Mean ($\bar{x}$)	14.0511	22.0159
Sample standard deviation (s)	3.5594	4.7971

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.

The histogram for the congruent word list (bin size = 10) shows that the data seems to be approximately normally distributed.



The same is true for the histogram for the answering times of the incongruent word list (same bin size).



The major differences we can observe from the plots are the distributions of answering times at the tails of the histogram. For congruent words we observe more entries of answer times in the 10s bucket, while for incongruent words we can observe the opposite, more entries in the 40s bucket. Furthermore, the centers of the distributions differ, as indicated by the descriptive statistics table in question 2.

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?

We are conducting a paired-samples t-test at α **= 0.05**, where:

- the mean difference, $\bar{d}$, is **7.9648**
- the standard deviation of differences**,** s_d**,** is **4.8648**
- the standard deviation of the differences **SE(**$\bar{d}$**)** is **0.9930**
- the t-statistic, **T**, is **8.0207** with **23 degrees of freedom**
- the **critical statistical value** is **2.069**

Given the t-statistic we **reject** the null hypotheses and conclude that there is statistically significant difference between response times of subjects when dealing with congruent and incongruent words.

The result of the test is in line with expectations, since I spent more time identifying incongruently labeled words myself when taking the test.

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!

In suspect that the perception of the color with our eyes and the comprehension of the text with our mind interfere when dealing with incongruently labeled words. As a consequence, the brain takes more time to resolve this conflict in relation to the standard situation when dealing with congruently labeled words.

A similar task causing this issue could be identifying geometric shapes (circles, cuboids, pyramids) when they are labeled incongruently.

Resources

- Udacity data analyst nanodegree documentation
- Bowerman et al (2014), Business Statistics in Practice 7th edition, McGraw-Hill
- Test assumptions, retrieved from: http://www.psychology.emory.edu/clinical/bliwise/Tutorials/TOM/meanstests/assump.htm