Udacity - Data analyst nanodegree
Statistics - Project submission

Gabor Galgocz

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

The independent variable in this test is the type of test, *congruent words* vs. *incongruent words*. The dependent variable is the time it takes to name 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.

We'll use μ_1 and μ_2 to represent the population mean for test 1 (congruent words) and test 2 (incongruent words).

The null hypothesis is that the population mean for test 1 and test 2 will be the same:

H_0: $\mu_2 - \mu_1 = 0$

In other words, the test type has no effect on the test results.

The alternative hypothesis is that the population means will be different for the two tests, this means that the speed of performing the task is affected by the fact whether the words are congruent or incongruent, though we don't know which one will be greater.

H_A: $\mu_2 - \mu_1 \neq 0$

In other words, we should be conducting a two-tailed test. Since the participants were asked to perform both tasks, we are talking about a dependent test, and since we don't know the population parameters, we are talking about a t-test.

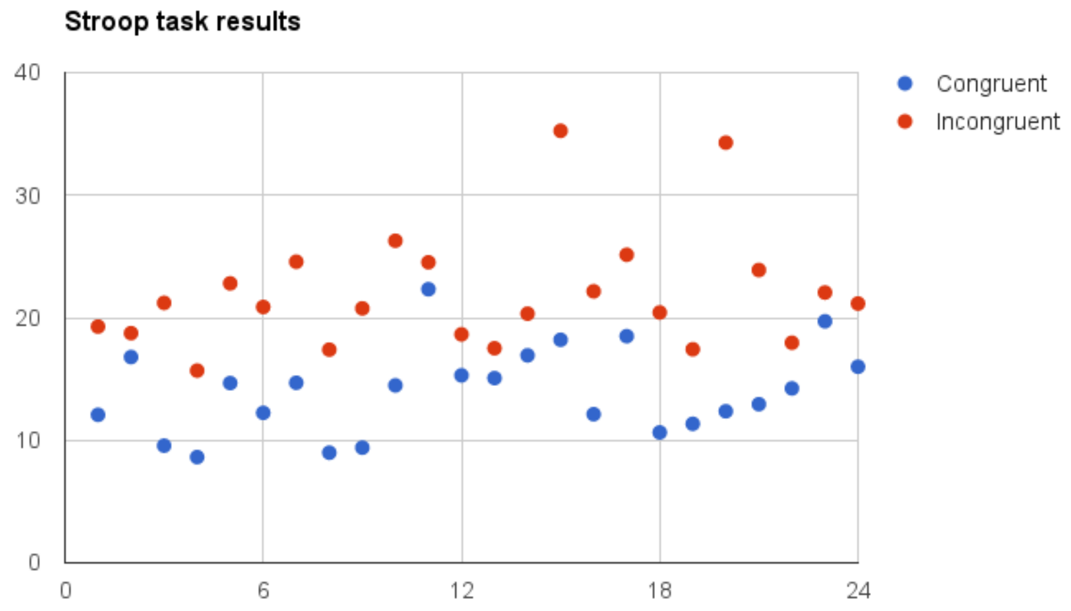
3. Report some descriptive statistics regarding this dataset. Include at least one measure of central tendency and at least one measure of variability.

Central tendency
For the sample with congruent words:
Mean: 14.05
Median: 14.36

For the sample with incongruent words:
Mean: 22.02
Median: 21.02

Variability
For the sample with congruent words:
Variance: 12.67
Standard deviation of the sample: 3.56

For the sample with incongruent words:
Variance: 23.01
Standard deviation of the sample: 4.80



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 best way to visualize distribution of sample values is to use a scatterplot. I used two colors to differentiate between the two samples, and the visualization made it clear that in all cases the task with incongruent words took more time to complete than the task with congruent words. Of course this can be seen just by looking at the raw numbers of the task results, but if the sample contains a lot more data points, the scatterplot is the easy way to understand the distribution of the results.

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?

alpha: 0.05
point estimate: 7.96
sample standard deviation of the differences: 4.86
t-statistic: 8.02
t-critical values: ± 2.069

The t-statistic falls in the critical region, we reject the null. This means the two samples are different.

Cohen’s d: 1.64
confidence interval: 95% CI = (5.91, 10.02)

We can conclude that it is indeed faster to finish the task with congruent words. This matches up with our expectations.