IKV1 homework - Hypotheses Testing on ETRA Challenge dataset

Martin Picek

December 2021

1 Introduction

In ETRA Challenge, 8 subjects were tested on various tasks including visual fixation, visual search, and visual exploration. Multiple stimuli were presented, for example, natural images, blank scenes, or a pictures scanned from Where's Waldo book. Based on the data, we came up with two hypotheses:

1. Pupil size is the same on blank and natural stimuli (during free viewing task).
2. Maximal pupil size is different when Waldo is found (during free viewing task).

2 Data format

The dataset is divided into two parts. The file `DataSummary.csv` where each trial has one row with clicks on the screen recorded. Then there is the `data` folder with a subfolder for each subject. For each trial of a subject, there is a file in the subject's subfolder.

3 Hypothesis 1: Pupil size is different on blank and natural stimuli (during the free viewing task).

For this experiment, we collected data of blank and natural stimuli separately (during the free viewing task). For each subject, we obtained the data from trials, concatenated the corresponding data with regard to stimulus and therefore obtained two data tables - one for natural stimulus, and one for blank stimulus.

Then we performed a statistical test testing whether the means of the two distributions are the same. We obtained a p-value for each subject.

3.1 What statistical test to use?

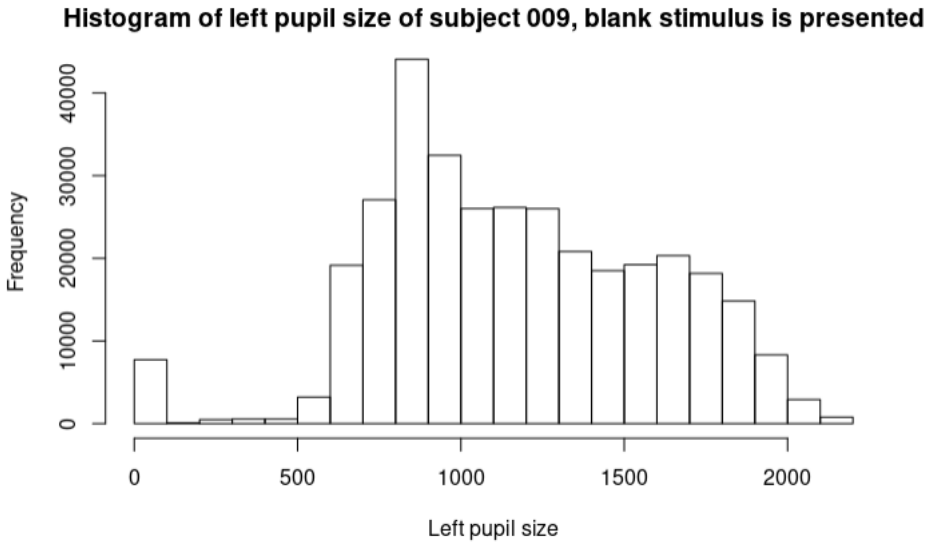
There are many statistical tests that we can choose from. One very widely used is called the T-Test. This test assumes that the two distributions follow the normal distribution, have unknown standard deviations and means μ_1 and μ_2. Then the H_0 states that $\mu_1 = \mu_2$, alternative H_1 states that $\mu_1 \neq \mu_2$.

The p-values obtained are the following:

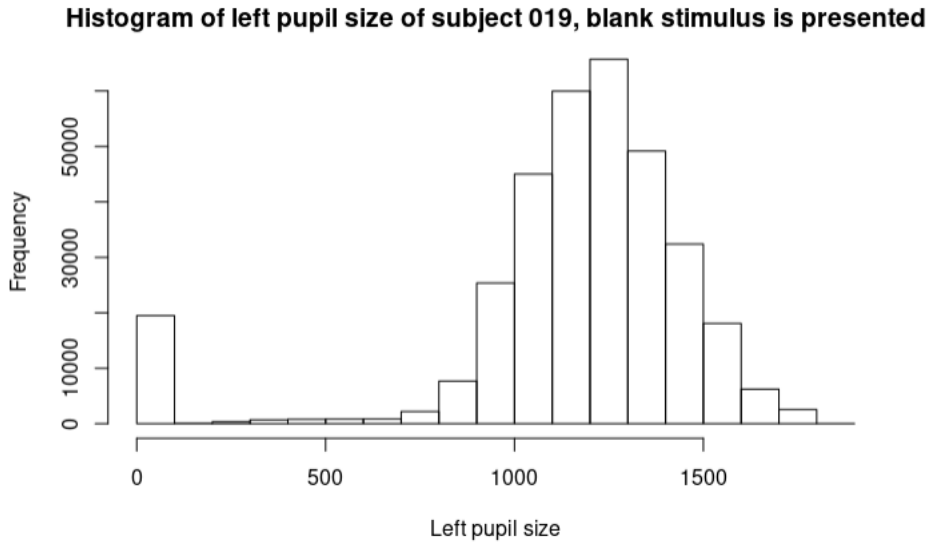
- Left pupil: (0 0.1071 0 0 0 0 0 0)
- Right pupil: (0 0 0 0 0 0 0 0)

What we found out is that if the significance level is 2 percent, 7 out of 8 tests rejected the null hypothesis (regarding the left pupil). That means that 7 out of 8 tests returned that the means are different (for the left pupil). On the other hand, p-values for the right pupils were all zeros - meaning all hypotheses are rejected.

However, this test assumes one important property - the distributions follow the normal distribution. But what if they do not? Here is a plot of left pupil size on blank stimuli of subject 009. Although no similarity test to normal distribution was performed, it is obvious that the distribution does not resemble the Gaussian curve.



On the other hand, the only test that has p-value larger than the significance level is very similar to the normal distribution. This is maybe the reason why this is the test with the higher p-value.



Owing to this observation, we performed the Wilcox rank-sum test. This test is a non-parametric analogy of the T-Test. While T-Test assumes that the distributions are normal, Wilcox

rank-sum test doesn't assume any specific distribution. It just tests the equality of means.

Let us see the resulting p-values obtained by Wilcox rank-sum test:

- Left pupil: (0 0 0 0 0 0 0 0)
- Right pupil: (0 0.3983 0 0 0 0 0 0)

Now we can see, that the test did not reject the H_0 hypotheses for the third subject.

3.2 Discussion of results

We discovered that the size of pupils differs between natural and blank stimuli. We came up with the following explanations:

1. **The images are different and therefore evoke different sizes of the pupil.** This might be explained by the fact that the subject sees reasonable objects on the natural stimulus and the eye is thus different. When the blank screen is presented, the eye is not really interested and the pupil size reflects the fact.
2. **The brightness of the stimuli are different**. This difference then changes the size of the pupils. This explanation is in our opinion the most credible one.
3. **The data is imperfect.** We have to take into account the possibility that the device is imperfect and hence our results are not precise.

4 Hypothesis 2: Maximal pupil size is different when Waldo is found (during Free Viewing task).

Working with data in this task was harder - the data is limited. Every trial has assigned a row of clicks on the screen. But there is no record of when the click was performed. We assume that when the subject clicks on the screen, the subject found the Waldo (and clicked on him).

Trials with no click are considered to be an unsuccessful trial of finding the Waldo - the subject did not find him (he/she did not click on him).

For each subject we divided the files of trials with our specific task into two sets:

1. Files with trials where Waldo was found
2. Files with trials where Waldo was not found

Each set contained multiple files. In fact, one subject had one set empty, we decided to discard the subject from this test as no comparison would be performed.

From each file in each set, we extracted 5 maximal pupil sizes due to our next assumption - the maximal pupil size is when the subject finds Waldo. This assumption had to be made since we do not have available times when the subject found Waldo.

These maximal pupil sizes were connected within each set into one data frame. Eventually, we obtained two data frames for each subject. One data frame containing maximal pupil sizes when Waldo was found, one data frame containing maximal pupil sizes when Waldo was not found.

The same thing was performed for the other eye.

4.1 What statistical test to use?

Now it was time to perform the test: H_0: The maximal pupil sizes are the same. H_1: The maximal pupil sizes are different.

Similarly to the first hypothesis above, we conducted 2 tests - the T-Test and Wilcox rank-sum test for each pupil. The reason why we did that is the same as in the first hypothesis.

Here are the p-values we obtained:

Eye	Test	subject1	subject2	subject3	subject4	subject5	subject6	subject7
Left	T-Test	4.565e-05	6.723e-14	2.828e-02	2.840e-02	1.910e-07	2.593e-15	2.746e-11
Right	T-Test	7.683e-02	8.843e-05	9.876e-02	1.546e-02	1.062e-05	1.947e-04	4.940e-11
Left	Wilcox	1.010e-04	1.003e-04	2.076e-02	5.226e-02	3.106e-07	6.759e-04	1.334e-08
Right	Wilcox	2.703e-02	3.886e-02	1.132e-01	1.684e-03	4.396e-06	1.003e-04	1.688e-06

The table is now more interesting than in the first hypothesis. We see that the significance level now plays an important role. Furthermore, Wilcox rank-sum test is almost always better for this task than the T-Test.

If we choose our significance level to be 1 percent, we would have to reject only 4 out of 7 subjects (considering right eye when Wilcox rank-sum test used as we want to be conservative in order to make as few Type I errors as possible). If the significance level is 0.01 percent, we would have to reject only 2 out of 7 subjects.

4.2 Discussion of results

The data are not really precise for this task and the results are also inconclusive. But as said, in order to make as few Type I errors as possible, the data can be considered to show that the alternative hypothesis cannot be accepted and H_0 hypothesis cannot be rejected.

5 Discussion and conclusion

We tested two hypotheses in total. The first one was successfully conducted, we rejected the H_0 hypothesis, and the alternative H_1 was accepted. Our p-values were significant (with just one sample deviating). Possible reasons for this behavior of the human eye were stated.

On the other hand, the second hypothesis testing was more difficult due to insufficient information provided in the dataset. It is a crucial fact to consider and we should regard this statistical test as inaccurate and inconclusive. However, the data indicates that the evidence is not significant enough to reject the H_0 hypothesis.