

HCI Assignment 3: Statistical Significance Test

Rohit Rawat

OBJECTIVE:

To understand the basics of statistical significance testing by analyzing given datasets and performing pairwise t-test [1] and Analysis of Variance (ANOVA [2]).

INTRODUCTION:

DATASETS:

Dataset1: It contains three columns: user, menu and time.

Independent Variable: menu

Dependent Variable: time

Dataset2: It contains four columns: user, menu, time and error.

Independent Variable: menu

Dependent Variable: time and error

As time and error changes on the basis of the type of menu, they are dependent variables. While menu type doesn't change and not dependent to any other attributes or feature, it is independent Variable.

TASK 1: BETWEEN-SUBJECTS DESIGN:

The first dataset contains **user** id, type of **menu** and task completion **time**. There was a total of 40 users, 10 each for a menu type. So, there were 4 groups of users and it was a **between-subjects** design.

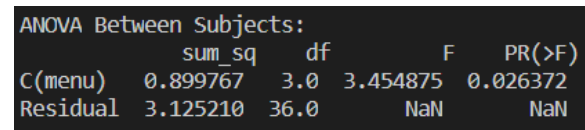
We designed the following null hypothesis for this experiment:

H₀ (null Hypothesis): It states the mean of the time taken for all the menu types would be same.

a. ANOVA analysis:

We first performed ANOVA between subjects analysis using **anova_lm()** of **statsmodels[3]** library of python.

Results: From figure1, we got **p-value = 0.026372**, which is less than 0.05. Thus, we can reject the H₀, null hypothesis and can say that means of time taken for all the different menu types, have distinct values i.e. it is statistically significant.



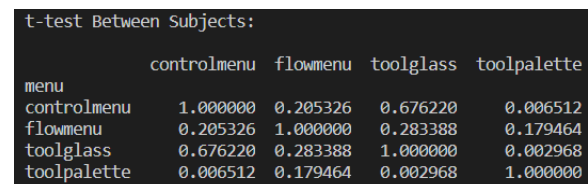
ANOVA Between Subjects:				
	sum_sq	df	F	PR(>F)
C(menu)	0.899767	3.0	3.454875	0.026372
Residual	3.125210	36.0	NaN	NaN

Figure 1. ANOVA Between-Subjects

b. Pairwise t-test analysis:

We performed pairwise t-test on the dataset using **ttest_ind()** function from **scipy.stats** library of python.

Results: From figure2, we can see in the cross grid for p-values that **controlmenu** vs **toolpalette** and **toolglass** vs **toolpalette** have statistically significant p-values as they are lesser than 0.05, i.e. **0.006512** and **0.002968**, respectively. Thus, it means they reject the null hypothesis and will have distinct mean values. Rest of the comparisons don't give statistically significant analysis.



t-test Between Subjects:				
menu	controlmenu	flowmenu	toolglass	toolpalette
controlmenu	1.000000	0.205326	0.676220	0.006512
flowmenu	0.205326	1.000000	0.283388	0.179464
toolglass	0.676220	0.283388	1.000000	0.002968
toolpalette	0.006512	0.179464	0.002968	1.000000

Figure 2. Pairwise t-test

c. Visualization for Between-Subjects Design:

- We depicted our data distribution using **boxplot** using **Seaborn** library of python in figure 3.
- We can observe that for the **toolglass** and **toolpalette**, the task time taken distribution for all the 10 subjects, were almost close to their median (center line of the box).
- From figure 3, we can also see that median value for the **controlmenu** is the lowest, thus it is better menu than the rest of the menus.

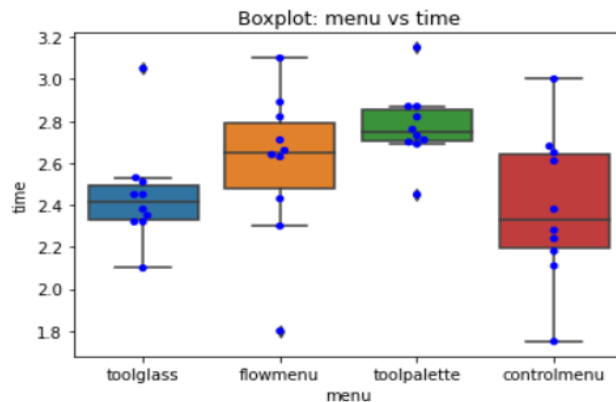


Figure 3. menu vs time

- From figure 4, we can notice the mean time value for **toolpalette (2.775s)** vs **controlmenu (2.388s)** and **toolpalette (2.775s)** vs **toolglass (2.446s)** have significant difference, which we have observed using **t-test analysis (p-value lesser than 0.05)**.

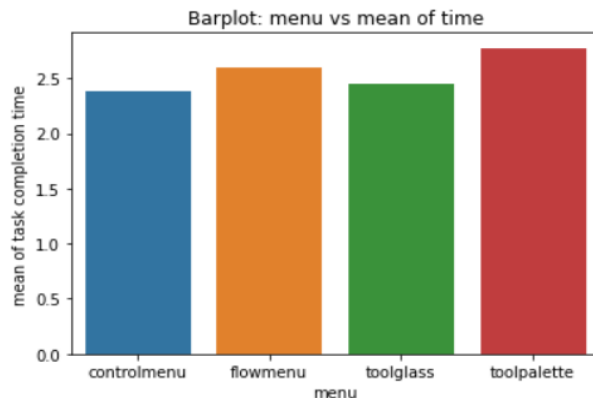


Figure 4. menu vs means of task completion time

- But **toolpalette (2.775s)** vs **flowmenu (2.598s)** has not have significant difference in the mean. Resultantly, p-value for this comparison from the t-test is **not statistically significant (p=0.179464>0.05)**.

TASK2: WITHIN-SUBJECTS DESIGN:

In the Second dataset we got **user id**, type of **menu**, task completion **time** and **error**. There is a total of 10 users, each user testing each menu type. It's a **within-subjects** design experiment.

Each user is going to perform all the 4 tasks; thus it is a repetitive measure design. We designed the following null hypothesis for this experiment:

H₀ (null Hypothesis): It states the mean of the time taken for all the different menu types would be same.

a. ANOVA analysis:

We performed ANOVA within subjects analysis using **AnovaRM()** function of **statsmodels** library of python.

Results: From figure3, we can see that p-value is **0.0001**, which is less than 0.005. Thus, we can reject the null hypothesis, H₀. And can say, that p-value is statistically significant for ANOVA repetitive measure.

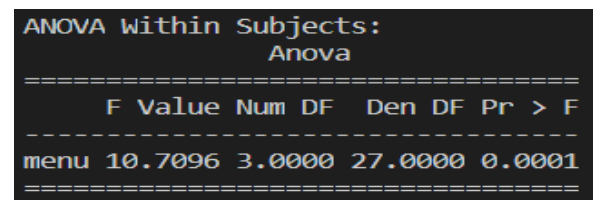


Figure 5. ANOVA Within-Subjects

b. Pairwise t-test analysis:

We performed pairwise t-test on the dataset using **ttest_rel()** function from **scipy.stats** library

of python. Unlike the Task 1, this experiment was dependent.

Results: From figure4, we can see in the cross grid for p-values that **controlmenu vs toolpalette**, **toolglass vs toolpalette** and **flowmenu vs toolpalette** have statistically significant p-values as they all are lesser than 0.05, i.e. **0.000508**, **0.001656** and **0.000439**, respectively. Thus, it means they reject the null hypothesis and will have distinct mean values. The **toolpalette** menu type has distinct mean value as compare to all the other menus. Rest of the comparisons don't give statistically significant analysis.

t-test Within Subjects:

	controlmenu	flowmenu	toolglass	toolpalette
menu				
controlmenu	NaN	0.845985	0.276323	0.000508
flowmenu	0.845985	NaN	0.076665	0.001656
toolglass	0.276323	0.076665	NaN	0.000439
toolpalette	0.000508	0.001656	0.000439	NaN

Figure 6. Pairwise t-test

c. Visualization for Within-Subjects Design:

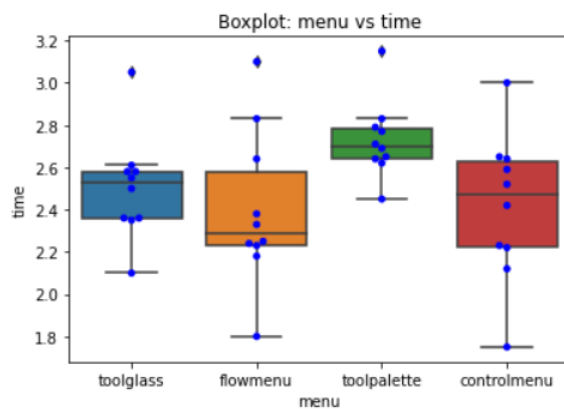


Figure 7. menu vs time

- We depicted our data distribution using **boxplot** using **Seaborn** library of python in figure 7 i.e. boxplot menu vs time.

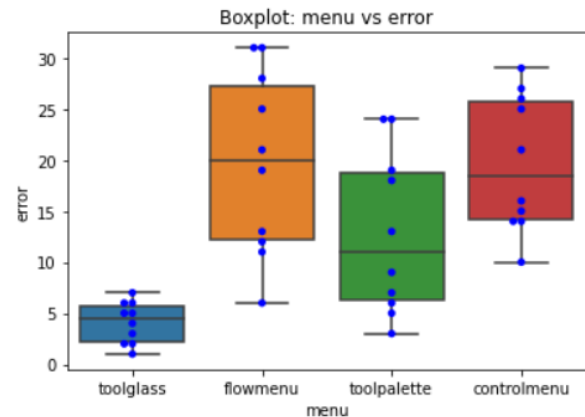


Figure 8. menu vs error

- Also, most of the points are distributed around the median, for all the different menus. But also, we can notice some **outliers** for the **toolglass**, **flowmenu** and **toolpalette**.
- Again, from figure 7, we can see that median value for the **flowmenu** is the lowest. Thus, it is faster than the rest of the menus.
- From figure 8, we can see the median error for the **toolglass** menu is the lowest and all the points are lying around the median. Thus, we can say that the **toolglass** menu task was accurately done and was clearer for the users to handle.

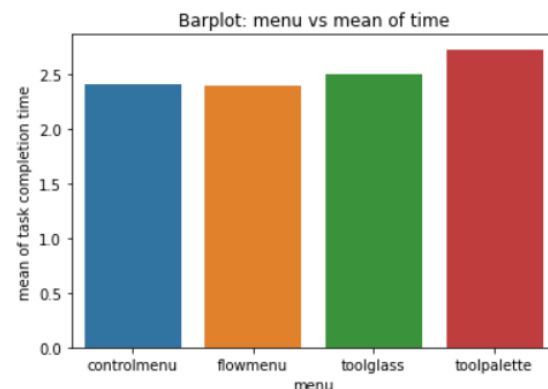


Figure 9. Menu vs mean of time taken

- From figure 9, we can notice that mean time taken for by the **toolpalette(2.730s)** menu is significantly higher than the rest of the menus(**toolglass: 2.504s**, **flowmenu: 2.398s**,

controlmenu: 2.414s). Resultantly, we got the p -values < 0.05 (statistically significant) while making comparisons of the **toolpalette** vs rest of the menus.

CONCLUSION:

We have successfully analyzed Between-subjects design and within-subject design experiment using ANOVA and t-test analysis methods on two different datasets. We have also visualized the datasets in order to understand the statistical significance of the two tests performed.

REFERENCES:

- [1]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html
- [2]<https://www.statsmodels.org/stable/anova.html>
- [3]<https://www.statsmodels.org/stable/index.html>