CS3106 HCI - ANOVA, Experiment Design Implementation

140012021

November 24, 2016

Contents

1 Introduction

1.1 Aims and Objectives

- To understand Analysis of Variance, how to determine, apply and interpret a F-Distribution value from an ANOVA.
- To understand how to design an experiment for comparing two menu styles, along with the factors influencing each, and the threats to validity.
- To implement and test the two menu styles sing your design.

2 Question 1 - One-Way ANOVA

2.1 Equations in use

$$SS = \sum_i (Xi - \bar{X})^2$$

$$SS_{error} = SS_A + SS_B$$

$$SS_{total} = SS_{A+B}$$

$$SS_{error} = SS_{total} - SS_{error}$$

$$df_{error} = nparticipants - mgroups$$

$$df_{effect} = mgroups - 1$$

$$MS_{error} = SS_{error}/df_{error}$$

$$MS_{effect} = SSeffect/df_{effect}$$

$$F - ratio = MS_{effect} - MS_{error}$$

2.2 Data Set 1 - Keyboard

	A	B			SS_A	SS_B	SS_{A+B}
	1	3			4	4	10
	2	4			1	1	4
	3	5			0	0	2
	4	6			1	1	4
	5	7			4	4	10
Mean	3	5		Sum	10	10	30
$Mean_{A+B}$	4						
SS_{error}	20						
SS_{total}	30						
SS_{effect}	10						
n Participants	10						
m Groups	2						
df_{error}	8						
df_{effect}	1						
MS_{error}	2.5						
MS_{effect}	10						
α Confidence Level	0.05						
F-ratio	4						
Critical value	5.31766						
Significant?	0						

Figure 1: Calculations of Means, Sum Square values, Degrees of freedom, Mean squares, F-ratio and Critical value for Keyboard data set.

- SS_{total}: 30
- SS_{error}: 20
- SS_{effect}: 10
- df_{effect}: 1
- df_{error}: 8
- F-ratio: 4
- F-Distribution value (critical value): 3.86118
- Significance level of reject null hypothesis: 0.085

Reporting: Keyboard A resulted in fewer errors than Keyboard B (3 vs. 5) We assumed these errors were normally distributed. Analysis of variance at a significance level of $\alpha = 0.05$ showed that this difference was not statistically significant. (Critical value > F-Ratio.)

2.3 Data Set 2 - Mouse

	C	D			SS_C	SS_D	SS_{C+D}
	2	8			16	16	50
	4	10			4	4	26
	6	12			0	0	18
	8	14			4	4	26
	10	16			16	16	50
Mean	6	12		SUM	40	40	170
$Mean_{C+D}$	9						
SS_{error}	80						
SS_{total}	170						
SS_{effect}	90						
***n* Participants**	10						
***m* Groups**	2						
df_{error}	8						
df_{effect}	1						
MS_{error}	10						
MS_{effect}	90						
α Confidence Level	0.05						
F-ratio	9						
Critical value	5.31766						
Significant?	1						

Figure 2: Calculations of Means, Sum Square values, Degrees of freedom, Mean squares, F-ratio and Critical value for Mouse data set.

- SS_{total}: 170
- SS_{error}: 80
- SS_{effect}: 90
- df_{effect}: 1
- df_{error}: 8
- F-ratio: 9
- F-Distribution value (critical value): 5.31776
- Significance level of reject null hypothesis: 0.05

Reporting: Mouse C resulted in fewer errors than Mouse D (6 vs. 12) We assumed these errors were normally distributed. Analysis of variance at a significance level of $\alpha = 0.05$ showed us that this difference was statistically significant. (Critical value < F-Ratio.)

2.4 Data Set 3 - Multi-Touch

	E	F	G			SS_E	SS_F	SS_G	SS_{E+F+G}
	28	30	51			61.7347	71.0408	88.898	545.075
	28	44	70			61.7347	31.0408	91.6122	916.361
	34	36	58			3.44898	5.89796	5.89796	369.694
	36	39	60			0.02041	0.32653	0.18367	342.027
	46	30	67			102.878	71.0408	43.1837	711.551
	38	46	62			4.59184	57.3265	2.46939	341.122
	41	44	55			26.449	31.0408	29.4694	117.98
Mean	35.8571	38.4286	60.4286		**SUM**	260.857	267.714	261.714	3343.81
$Mean_{E+F+G}$	44.9048								
SS_{error}	790.286								
SS_{total}	3343.81								
SS_{effect}	2553.52								
***n* Participants**	21								
***m* Groups**	3								
df_{error}	18								
df_{effect}	2								
MS_{error}	43.9048								
MS_{effect}	1276.76								
α Confidence Level	0.05								
F-ratio	29.0803								
Critical value	3.55456								
Significant?	1								

Figure 3: Calculations of Means, Sum Square values, Degrees of freedom, Mean squares, F-ratio and Critical value for MT data set.

- SS_{total}: 3343.81
- SS_{error}: 790.286
- SS_{effect}: 2553.52
- df_{effect}: 2
- df_{error}: 18
- F-ratio: 29.0803
- F-Distribution value (critical value): 3.55456
- Significance level of reject null hypothesis: 0.05

Reporting: MT E resulted in fewer errors than MT F and both resulted in fewer errors than MT G (35.8 vs. 38.4, vs. 60.4 respectively) We assumed these errors were normally distributed. Analysis of variance at significance level of $\alpha = 0.05$ showed us that these differences were statistically significant.

2.5 Evaluation

As with all One-Way ANOVA studies just because it says not to accept the null hypothesis doesn't mean you can definitely reject it. If we are using a confidence level of 0.05 and it says to reject the null hypothesis it means we only have a 95% chance of rejecting it successfully. Data could be collected in future studies to accept the null hypothesis. 0.05 is seen as a golden value to do the study with and is the usually used as it is very unlikely to ever be 100% rejecting a null hypothesis. The reason degrees of freedoms are included in calculating the critical value is because we are using a sample not a population. And by using df we are reducing the effects of using a sample. Having more people take part in a study lowers these numbers making the results clearer and makes us more confident to reject the null hypothesis. By doing using the Between Subjects it stops there being a transfer of skills between the two options, but it also means that if someone makes an error it has a greater bearing on the overall result of the ANOVA.

You can see from the averages on each set what was going to be significant and not. For Keyboard there is only 25% in the errors which is close for so few errors while for the Mouse it is 50% and it is a similar value again for the MT. These can gve you an early indication on which set can be rejected. Also there are more people on each device for the MT than the other devices. This means that the df are higher meaning that if someone makes an error it will effect the results in a less significant way.

For the Keyboard we get that it is not significant for an acceptance value of 0.05 but we do for 0.085. This means we can accept the null hypothesis with a confidence level of 91.5%. This is slightly lower than you would like to be able to accept it. For the Mouse we get that is is significant for an acceptance value of 0.05. This means we can accept the null hypothesis with a confidence level of 95%. This is a good value to accept it and move on too the next stage of research. For the Multi-touch system we get that is is significant for an acceptance value of 0.05. This means we can accept the null hypothesis with a confidence level of 95%. This is a good value to accept it and move on too the next stage of research.

3 Question 2 - Menu Comparison and Experimental Design

3.1 Factors when Choosing Menu

3.1.1 Spacial issues

The main resource a menu takes up on a computer is part of the screen capacity. This is generally a fixed size and depending on that size, each menu could be used in a useful manner. If the number of items on the list could go up exponentially then a linear menu would be better suited to a limited screen size, by normally having a scroll feature in the menu. While a pie menu "become polynomially larger than linear menus in both height and width with increased item size and number of items." [1] So if there is a limited screen size with a chance of the list of items increasing a linear menu would be preferred.

3.1.2 Items in the Menu

Menus are there to display many options to a user. These options come in different categories that suit different types of menus. A linear menu suits items of a sequential nature, such a 1,2,3 or a,b,c due to the fact that a linear displays items one after another one on the screen. A Pie menu struggles with these sort of items due to there often being no logical place to start these in comparison to linear menus. Alternatively pie menus are much better suited to compass directions, angular degrees, diametrically opposed or orthogonal function names. A user will no instinctively where North on a compass is and go there on a pie menu, which will be quicker than finding North half way down a sequential list on a linear menu.

3.1.3 User Input and Device

The speed at which each menu and can be navigated is down to the input methods and the device it is being displayed on. If you were designing a menu for a smart watch with a circular face a pie menu would be suited for the device. Not only would the menu use the space very effectively, while also being able to effectively display time. Another feature of pie menus is making selection about how far you go in a direction. i.e By going further up with the input you could get another input. This is possible and shown in WatchMI.[2] A linear menu benefits when the user has a keyboard to make entries. In many linear menus there are keyboard shortcuts to reach different parts of the menu quickly. So depending on the device being used each menu can be superior.

3.1.4 Speed of Input

Another important factor that needs to be considered when choosing a menu is the speed at which you want to access items in the menus. In a linear menu the first item in the list can be defaulted to be the automatic first option. So each other item in the list will take longer to get too, even using techniques mentioned above will usually still require some more movement of the cursor by the user. While in pie menus due the cursor starting in the middle of the screen it has the same distance to travel to each item, making it on average much quicker than a linear menu when trying to access a random item. So if the main thing you care about when implementing a menu is the speed to each item a pie menu is the much better option.

3.2 Experimental Design

Hypothesis - Pie Menus will be quicker than Linear Menus for radial type data and data with no type at all. While linear Menus will be quicker for sequential data. Pie Menus will also return a higher accuracy than Linear Menus.

Independent Variables - Size of Menus. Items on Lists. The Menu type.

Dependant Variables- Time taken for item to be selected and Accuracy of selection.

Participants - All participants in the study are not in the 3106 HCI module. This is to stop people who have experience with a similar experiment experience taking part in the study which might skew results. In total 6 would take part in the between subject study selecting 15 items. Doing this once for each item type (seq,radial,random).

Apparatus - A laptop, mouse, mouse mat and program. The mouse and mouse mat were used to make selections as easy a possible for participants as some might have issues with track pads.

Study Procedure and Set Up - each participant completed the study in the same control environment, while I supervised them carry out each trial. Each participant was shown the how the menus work, then were allowed to make a five selections. This was to make sure the timing was accurate for the first few runs as if the practise wasn't allowed users might need to learn it first.

Extra Info - The order at which the items in each trial changed order. This was to reduce skill learning effects than could be generated if they started to memorize orders. The three different sets of trial data were also given in random orders due to them being different (numbers or words). This was again to reduce possible skill transference. Finally the fact that the study was a Between-Subject study there wouldn't be skill transfer between the menus as they would only be testing one.

4 Question 3 - Empirically Comparison of Menus.

4.1 Overview and Disclaimer

The empirical comparison of the two menus will follow the experiment design mentioned in Question 2. The results will then be put several One-Way ANOVAs to see if we can reject the null hypothesis to show there is a difference in the menus for depending on the different types of items in the menus.

There is a documented java bug in the linear menu code. One of the mouse events will only ever return false when used on Windows OS. This means that the menu can only be used in Linux and Mac OS environments.

4.2 One-Way ANOVA Results

4.2.1 Comparison of Menus for Radial Type Items

4.3 Data Set 1 - Radial Items

	Pie Menu	Linear Menu		$SS_{PieMenu}$	$SS_{LinearMen}$	$SS_{PieMenu+LinearMenu}$
	1.432044218	1.76157		0.00175	0.06419	0.0766560635
	1.409956665	1.422647		0.00409	0.00732	0.0112598281
	1.579732	1.340445		0.0112	0.02815	0.0305483221
Mean	1.473910961	1.5082206667	**Sum**	0.01704	0.09966	0.1184642137
$Mean_{PieMenu+LinearMenu}$	1.4910658138					

SS_{error}	0.1167
SS_{total}	0.11846
SS_{effect}	0.00177
***n* Participants**	6
***m* Groups**	2
df_{error}	4
df_{effect}	1
MS_{error}	0.02917
MS_{effect}	0.00177
α Confidence Level	0.05
F-ratio	0.06052
Critical value	7.70865
Significant?	0

Figure 4: Calculations of Means, Sum Square values, Degrees of freedom, Mean squares, F-ratio and Critical value for Radial data set.

- SS_{total}: 0.11846
- SS_{error}: 0.1167
- SS_{effect}: 0.00177
- df_{effect}: 1
- df_{error}: 4
- F-ratio: 0.06052
- F-Distribution value (critical value): 7.70865
- Significance level of reject null hypothesis: 0.85

We have a 15% level of confidence for rejecting the null hypothesis.

4.3.1 Data Set 2 - Sequential Items

		Pie Menu	Linear Menu			$SS_{PieMenu}$	$SS_{LinearMenu}$	$SS_{PieMenu+LinearMenu}$
		1.361949	1.38934			0.1493961643	0.0225165	0.0283325086
		1.96099	1.164			0.0451660255	0.00566783	0.3270136644
		1.922462	1.164515			0.03027426	0.00559055	0.2921646277
Mean		1.748467	1.239285		Sum	0.2248364499	0.03377489	0.6475108007
$Mean_{PieMenu+LinearMenu}$		1.493876						
SS_{error}	0.258611337							
SS_{total}	0.647510801							
SS_{effect}	0.388899464							
n Participants	6							
m Groups	2							
df_{error}	4							
df_{effect}	1							
MS_{error}	0.064652834							
MS_{effect}	0.388899464							
α Confidence Level	0.05							
F-ratio	6.015195902							
Critical value	7.708647422							
Significant?	0							

Figure 5: Calculations of Means, Sum Square values, Degrees of freedom, Mean squares, F-ratio and Critical value for Sequential data set.

- SS_{total}: 0.64751
- SS_{error}: 0.258611
- SS_{effect}: 0.38889
- df_{effect}: 1
- df_{error}: 4
- F-ratio: 6.01519
- F-Distribution value (critical value): 7.70865
- Significance level of reject null hypothesis: 0.1

We have a 90% level of confidence for rejecting the null hypothesis.

4.3.2 Data Set 3 - Random Items

		Pie Menu	Linear Menu		$SS_{PieMenu}$	$SS_{LinearMenu}$	$SS_{PieMenu+LinearMenu}$
		1.45567	1.457783		0.0054902	0.0095609936	0.0002827141
		1.713403	1.317075		0.0337227	0.0018427846	0.0884375469
		1.420224	1.30515		0.0119994	0.003008815	0.0201337612
Mean		1.5297656667	1.3600026667	**Sum**	0.0512122	0.0144125932	0.1088540222
$Mean_{PieMenu+LinearMenu}$		1.4448841667					
SS_{error}	0.06562						
SS_{total}	0.10885						
SS_{effect}	0.04323						
***n* Participants**	6						
***m* Groups**	2						
df_{error}	4						
df_{effect}	1						
MS_{error}	0.01641						
MS_{effect}	0.04323						
α Confidence Level	0.05						
F-ratio	2.63493						
Critical value	7.70865						
Significant?	0						

Figure 6: Calculations of Means, Sum Square values, Degrees of freedom, Mean squares, F-ratio and Critical value for Random data set.

- SS_{total}: 0.10885
- SS_{error}: 0.06562
- SS_{effect}: 0.04323
- df_{effect}: 1
- df_{error}: 4
- F-ratio: 2.63493
- F-Distribution value (critical value): 7.70865
- Significance level of reject null hypothesis: 0.2

We have a 80% level of confidence for rejecting the null hypothesis.

4.4 Experiment Report

The original hypothesis for the experiment was to show that pie menus would be better for selecting items such as North, South, East, West since they are suited to being displayed in a circular fashion. While linear menus would be better for sequential items such as 1,2,3,4. And for random items such as Dog,Red,House a pie Menu would be faster because it has the same distance to travel for each selection but a linear menu has different distances, which is harder for a user to control.

For the first part of the study the acceptance value needed was 0.85. This gives us a very low confidence level for rejecting a null hypothesis. And this is correct for the data. The average time for each menu on the data set is nearly identical so there is no real difference in the menu timings meaning we can't reject the null hypothesis meaning that no menu is quicker than the other.

For the second part of the study the acceptance value needed was 0.1. This gives us a high confidence level for rejecting the null hypothesis that there is not a difference in the menus. And this seems correct for the data. The average time for the linear menu was nearly 0.5 seconds quicker than the pie menu for sequential items, which is significant for timings of under two seconds. This level of confidence means that there is a good likely hood that liner menus are quicker than pie menus for sequential items, matching our studies hypothesis.

For the third part of the study the acceptance value was 0.2. This gives us a fairly high confidence level for rejecting the null hypothesis that there is a difference in the menus. However it is the opposite of our studies hypothesis. It is the linear menu that has a lower average time than the pie menu. We have a fairly good confidence level for rejecting the null hypothesis that there is no difference in the menus. So there is a difference in the menus, just that liner menus are quicker for random items than pie menus, the opposite of what we thought.

4.4.1 Reflection

After looking back on the results of the experiments and speaking the participants after the study I believe I know why the studies hypothesis was wrong in parts. Every user that took part in the study had been using linear menus they're entire lives. So would all be experts in using them. But no one had ever used pie menus before before taking part in the study. I think if the subjects had experience in pie menus, then pie menus would be quicker for radial items, considering they are currently very equal. They would of also maybe become faster using pie for random items making them faster than linear items too. A longer study would be needed to test this new hypothesis.

4.4.2 Error Rates

In total there were 16 errors in the pie menu system that were evenly distributed across the different items types while there was just one error in the lineal menu across all data types. This was even after setting a region in the pie menu where users couldn't click in the middle where the angles were smaller, to keep errors down. I believe the reasons for these results are the same as in the reflection section.

5 Within-Subjects Comparison

In a Within-Subjects comparison each participant is exposed to all conditions. In this case they use both the menus not just one of the menus like in the Between-Subject comparison done earlier. The advantage of carrying out one of these studies is that Variance is controlled within the participant, which means you need fewer participants than a Between-subject comparison. However they're are disadvantages as well. More care is required to get the counter-balancing of start condition right to avoid skill transfer effects. There is also a risk of asymmetrical skill transfer where no matter what order you do the menus in you will be better at the second one. To try and stop this for this study, participants did the study in different orders. with the same amount doing linear first as the amount doing pie first. The items also appeared at a random order that users couldn't predict in the menus to try and balance conditions more.

For this study I was just comparing the two menus on random items and not multiple sets of items.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Menus	Sphericity Assumed	39.088	1	39.088	2.817	.154
	Greenhouse-Geisser	39.088	1.000	39.088	2.817	.154
	Huynh-Feldt	39.088	1.000	39.088	2.817	.154
	Lower-bound	39.088	1.000	39.088	2.817	.154

Figure 7: Significance values for repeated measures.

The significance value in the image is .154. This means we have a confidence level of 84.6% for rejecting that there is a difference in the menus. This is a higher confidence level than the One-Way ANOVA gave for the menus. This means that the Within-Subjects comparison may of given better results. It could of also gave back a higher number if they're was errors in the Between-Subject study. These would effect it more than a Within-Subject comparison would due to the reasons described above.

6 Submission Contents

- Location of Video: https://caf8.host.cs.st-andrews.ac.uk
- Report
- Source Code for Menus
- Repeated Measures File

7 Running Code

The Pie Menu has been created in processing so that will need to be downloaded to run the menu on and as mentioned previously the linear menu doesn't work on Windows due to a documented java bug.

8 References

1. An Empirical Comparison of Pie vs Linear Menus:
   https://www.cs.umd.edu/ ben/papers/Callahan1988empirical.pdf
2. WatchMI: http://sachi.cs.st-andrews.ac.uk/research/interaction/wearable/