IS5101 Practical 4: Data Analysis

Usability of two map-based fault reporting systems

#160022011
School of Computer Science
University of St Andrews

1 Introduction

In this practical a data set which was produced during a usability experiment was analysed in order to verify a number of underlying hypotheses. The conducted experiment involved a comparison between two map-based road incident reporting systems that provide a web-based interface. The hypotheses are being verified by analyzing the gained data from the experiment in a statistically valid way which involves techniques such as correlations, regressions and statistical tests.

The data analysis is structured as follows: Section 2 specifies the research hypotheses which the experiment was designed for and moreover introduces additional hypotheses about the experiment. Section 3 provides a concise and comprehensive description of the data set while section 4 offers essential information about the participants at a glance. Finally, in section 5 the established hypotheses are analysed and verified by statistically valid tests. Additionally, correlations between the different variables are analysed in this section and the influence of the within-subject design of the experiment on the results is examined.

2 Hypotheses

In order to compare the two fault-reporting systems in terms of their usability the following hypotheses shall be proven in this data analysis:

> H_{0_i}: *There is no significant difference between the times needed to locate faults using MapA and MapB.*
>
> H_{a_i}: *There is a significant difference between the times needed to locate faults using MapA and MapB.*
>
> H_{0_ii}: *There is no significant difference between the times needed to type fault descriptions using MapA and MapB.*
>
> H_{a_ii}: *There is a significant difference between the times needed to type fault descriptions using MapA and MapB.*

As common standard usability definitions, such as ISO 9241-11 (1998) ascribe usability not only properties such as time efficiency but also accuracy, following supplementary hypotheses are added to be proven using the available data from the data set:

> H_{0_iii}: *There is no significant difference between the accuracy of locating faults in MapA and MapB.*
>
> H_{a_iii}: *There is a significant difference between the accuracy of locating faults in MapA and MapB.*
>
> H_{0_iv}: *There is no significant difference between the accuracy of typing fault descriptions in MapA and MapB.*
>
> H_{a_iv}: *There is a significant difference between the accuracy of typing fault descriptions in MapA and MapB.*

3 Essential information about the data set

In total, 20 hours of data were collected during the experiment. In average, each participant reported 169.65 faults during the tests altogether. The analysed data set comprises a table of 14 columns × 3,393 rows.

As the participants were asked to report faults during the experiment, each row maps each combination of *pid*, which identifies the participant and *fid*, which identifies the fault, to the data collected during the experiment. This data comprises the longitudes, latitudes and text entered by the participants as well as the used map system and the finding and typing times (*f.time* and *t.time*). Statistical findings about the participants are described in the next section.

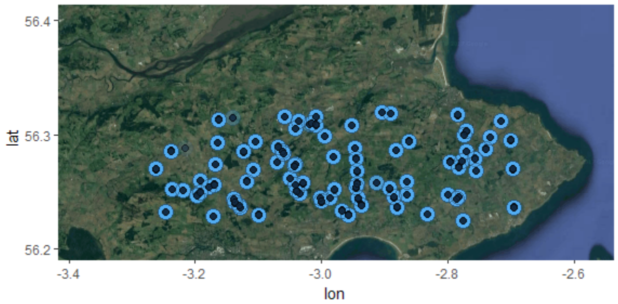


Figure 1: Map of the faults' locations; the small black circles depict the correct locations of the scattered faults; the bigger blue circles depict the reported fault coordinates gained during the experiment by the participants; the more dense the blue is, the more data points exist.

4 Participant Information

In total, 20 participants were recruited for the experiment. Following information was gained that describes characteristics of the participants in the data set:

The participants' ages ranged between 22 and 43 ($mean = 33.05$, $median = 33.5$, $sd = 5.80$, *see Figure 2*). Nine persons were male and ten persons were female (*See Figure 3*). One person in the data set did not specify any gender. In total, the twenty participants originated from seven different countries. Most participants were from Scotland and the US while only one person was from Spain (*See Figure 4*).

By conducting a first analysis at the times and accuracies per country, it was found that the times are close together. Only the typing error per country varied clearly as can be seen in *Figure 5*. A statistical analysis of the times and errors which includes significance testing is conducted in the next section.

Also the relations between gender and times, and respectively accuracy were analysed. In terms of finding faults, male and female participants were almost equally fast and accurate. As for typing descriptions both groups were equally fast, but male participants produced more typing errors than female participants with MapA (*See Figure 6*).

5 Analysis

This section discusses the statistical findings in the data set and verifies the hypotheses which were defined initially. In the following sub-sections, the different hypotheses are being verified in statistically valid ways. Additionally, this section offers relevant information for the researchers such as correlations between the variables and an analysis about the impact of chosen within-subject design on the result.

It was observed that with MapA participants reported 1713 faults in total while with MapB they only reported 1680 faults. Hence, each participant reported 85.65 faults ($sd = 1.73$) in average with MapA and only 84.00 faults ($sd = 1.52$) with MapB. These findings provide a first hint that working with MapA is faster than working with MapB which is proved in the following two sub-sections.

5.1 Time needed to locate faults

The times needed to locate a fault can be assumed to be normally distributed (See Figure 7). Therefore, the mean and the standard deviation are descriptors which characterise the times very accurately and enable comparability.

It was found that the mean localisation times vary significantly between MapA and MapB. While with MapB participants needed 20.13 seconds in average to locate a fault ($sd = 2.69$), with MapA participants only needed 19.69 seconds ($sd = 2.87$), which means that participants using MapA located faults faster than with MapB during the test. The significance of the means could be confirmed by a two-tailed Welch t-test ($\alpha = 0.05$, $p = 0.000004$), which is designed to test distributions with unequal variances, and which is chosen as no mean could be expected to be greater a priori. Therefore, the Null Hypothesis H_{0_i} can be discarded in favour of the alternative Hypothesis H_{a_i}.

5.2 Time needed to type fault descriptions

Also the times needed to type in fault descriptions into the systems can be assumed to be normally distributed (See Figure 8). By comparing the means it was found that, also for text entry times, the participants typed faster when using MapA. While with MapB participants required 1.84 seconds in average ($sd = 0.20$) per description, with MapA they only required 1.80 seconds ($sd = 0.19$), which makes MapA faster than MapB. The significance of these findings was confirmed by a two-tailed Welch t-test ($\alpha = 0.05$, $p = 3.568 \times 10^{-9}$), as no mean was expected to be greater. Therefore, the Null Hypothesis H_{0_ii} can be discarded in favour of the alternative Hypothesis H_{a_ii}.

5.3 Localisation accuracy

In order to compare the accuracy of the two systems for locating faults, the localisation errors for all fault reports were estimated. The localisation error of an entered location is the deviation between the correct location coordinates of the fault and the coordinates which were entered into the systems by each participant during the experiment. This deviation can be calculated by using the *Euclidean Distance* between the correct and reported coordinates.

$$Err_{loc} = \sqrt{(lat_{cor} - lat)^2 + (long_{cor} - long)^2} \qquad (1)$$

Applying this formula on the data set, the estimated mean localisation error was determined as 6.61×10^{-5} for MapA and 6.23×10^{-5} for MapB, which makes MapB's error smaller. In order to determine significant difference between the mean localisation error of MapA and MapB a non-parametric Mann–Whitney–Wilcoxon (MWW) test was conducted instead of a t-test, as the distributions cannot be assumed to be normal (See Figure 9).

The MWW-test ($\alpha = 0.05$, $p = 0.04072$) accounted for a significant difference between the means. Therefore, the Null Hypothesis H_{0_iii} can be discarded in favour of the alternative Hypothesis H_{a_iii}.

5.4 Typing accuracy

As the participants had to type in fault descriptions for each fault, the two systems are being compared in terms of their typing accuracies by comparing the typing errors made. The relative typing error for each description was calculated by dividing the *Levenshtein minimum edit distance* between the given fault description and the typed in fault description by the number

$mean_{MapA}$	sd_{MapA}	$mean_{MapB}$	sd_{MapB}	$p_{0.95}$
		f.time		
19.69	2.87	20.13	0.05	0.000004
		t.time		
1.80	0.19	1.84	0.20	3.568×10^{-9}
		loc.err		
6.61×10^{-5}	5.17×10^{-05}	6.23×10^{-5}	$4.90 \times 10-05$	0.04072
		typ.err		
0.0957	0.092	0.0947	0.094	0.632

Table 1: The table compares the four variables *f.time*, *t.time*, *localisation error* and *typing error* for the systems MapA and MapB in terms of mean, standard deviation and significance.

of characters in the correct description.

$$Err_{typ} = \frac{Dist_{Levenshtein}}{StrLen_{Descr}} \tag{2}$$

By applying the formula on the data set, the mean typing error was determined as 9.57% for MapA and only 9.47% for MapB, but the median was found to be identical for both errors ($median = 0.09$). As in section 5.3, the distribution of the typing error is non-normal (*See Figure 10*) the non-parametric MWW-test was conducted in order to verify, whether there is a significant difference between the two typing errors.

The MWW-test ($\alpha = 0.05$) resulted in a p-value of $p = 0.632$ which implies that the error rates of the two systems are not significantly different. Therefore, the Null-Hypothesis may be kept as the alternative Hypothesis is discarded and it can be concluded that the two systems are indifferent in terms of their typing errors.

5.5 Correlation between variables

In this section, several relations between the different variables have been examined in order to determine relational patterns between them. In order to determine these patterns, correlations between the variables were calculated and analysed.

First it was attempted to verify the intuitive assumption that *f.time* behaves similarly to *t.time* for the participants, which means that if a participant took a long time to find a fault in average the participant was likely to take a long time for typing in the description as well and vice versa. The averages of *f.time* and *t.time* per participant were plotted and their correlation was calculated (*See Figure 11*). A correlation coefficient of $R = 0.691$ indicates a clear correlation between the two variables. It could be confirmed as significant with a correlation test for Pearson's product-moment correlations ($\alpha = 0.05$, $p = 0.0007$).

Negative correlation ($R = -0.566$) was found between the average localisation error and the typing error per participant, which means that if a participant located the fault more precisely, the participant was more likely to have typing errors in the descriptions and vice versa. The significance of this finding could be confirmed with a correlation test ($\alpha = 0.05$, $p = 0.009$).

However, no significant correlation was found between times and errors. For the localisation error and the average *f.time* per person, the correlation coefficient ($R = -0.271$) indicates a slightly negative correlation but the correlation test turned out to be negative ($\alpha = 0.05$, $p = 0.2477$). For the typing error and the average *t.time* per person, the correlation coefficient was insignificantly small ($R = -0.051$, $p = 0.8325$).

Additionally, the correlation between the participants' ages and the times as well as the errors was analysed (*See Figure 12*). It was found that the age correlates positively with the typing error ($R = 0.597$, $p = 0.0008$) and negatively with the localisation error ($R = -0.750$, $p = 0.4.348 \times 10^{-06}$). It can be concluded that younger people were likely not to adhere to the exact position of a fault location as precisely as the older participants but in return seemed to have better typing capabilities. The times on the other side do not correlate with the participants' ages.

5.6 Influence of the experiment design

In this section, it is analysed what influence the choice of the experimental design has on the times and errors. Charness, Gneezy, and Kuhn (2012) discourage using within-subject designs as they are more susceptible for biases than the more conservative between-subject design. As it was decided to use a within-subject design for the experiment it shall be proven if the fact that both groups of participants use both system in a sequential order, testing the first system affects their behaviour when testing the second system. It was expected that participants may transfer knowledge from the first system to the second one, which might result into biased times and errors.

By comparing the average times and errors for each system between the participants who tested the system first and the participants whoe tested the system second, it was found that there is no significant difference in either of the time and error variables, apart from the typing error for MapB (*Table 2 & Figure 13*). The MWW-test indicates a significant difference of the typing error distributions for MapB: The participants who already tested MapA made less errors with MapB ($mean = 0.085$, $median = 0.0625$) than the participants who tested MapB first ($mean = 0.104$,

[1] Surprisingly, the test did not indicate significance for the typing error in MapA in terms of trained and untrained participants,

median $= 0.125)$[1]. It can be concluded that the finding that participants who tested MapB after MapA had a better accuracy than participants who tested MapB first, provides an indication that participants were influenced positively by MapA in terms of their typing accuracy while testing MapB.

6 Conclusion

In the context of this data analysis it has been proven that the participants using MapA were significantly faster in terms of localising faults and writing fault descriptions. However, by using MapB a higher accuracy in terms of localising fault locations could be achieved. There was no significant difference between the two systems in terms of typing errors.

Furthermore, a great deal of information could be extracted from the data set, which is interesting and important to the researchers who performed the experiment, such as participant information, correlation between different variables and impacts of the within-subject design on the experiment.

References

ISO. (1998). Ergonomic requirements for office work with visual display terminals (vdts) part 11: Guidance on usability.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. doi:10.1016/j.jebo.2011.08.009

Words in text: 2089

A Tables

Table 2: For both systems the table compares two groups of participants for each of the variables f.time, t.time, localisation error and typing error. The "untrained" group used a particular system first while the "trained" group used the other system before. The only significant difference between those groups was found in the typing error of MapB. The p-values were calculated with t-tests for the times and MWW-tests for the errors.

Variable	$\mu_{MapA}^{trained}$	$\mu_{MapA}^{untrained}$	p_{MapA}	$\mu_{MapB}^{trained}$	$\mu_{MapB}^{untrained}$	p_{MapB}
f.time	19.67302	19.71244	0.762	20.18522	20.08339	0.4681
Err_{loc}	6.657028×10^{-05}	6.569254×10^{-05}	0.9697	6.129206×10^{-05}	6.327433×10^{-05}	0.5914
t.time	1.796697	1.809406	0.1674	1.843277	1.842300	0.9205
Err_{typ}	0.1034543	0.0878579	0.138	0.08546874	0.1039279	**0.01758**

even though the average errors are opposed to the ones in section 5.6 (MapA tested as first system: *mean* $= 0.103$, *median* $= 0.125$; MapAtested as second system: *mean* $= 0.0879$, *median* $= 0.0625$).

B Figures

Figure 2: The age distribution of the participants depicts an age range between 22 and 43 ($mean = 33.05$, $median = 33.5$, $sd = 5.80$).

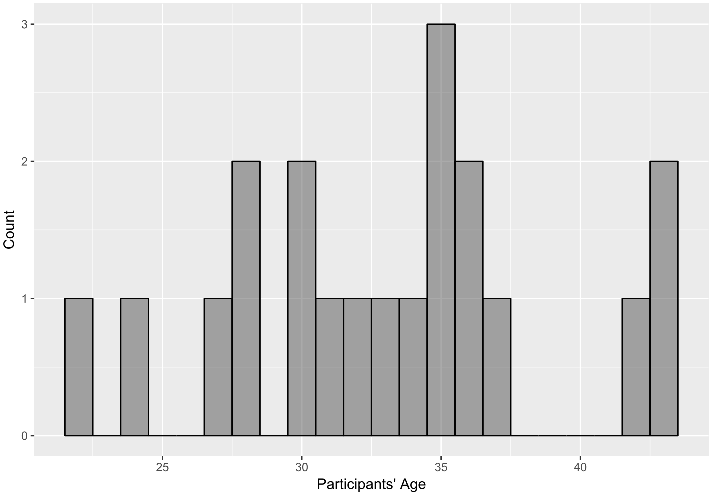


Figure 3: Genders are practically equally distributed. Only one participant did not specify the gender.

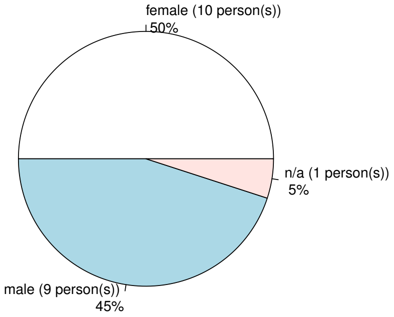


Figure 4: Participants originate from seven different countries. The majority of the participants is Scottish or American; only one participant is from Spain.

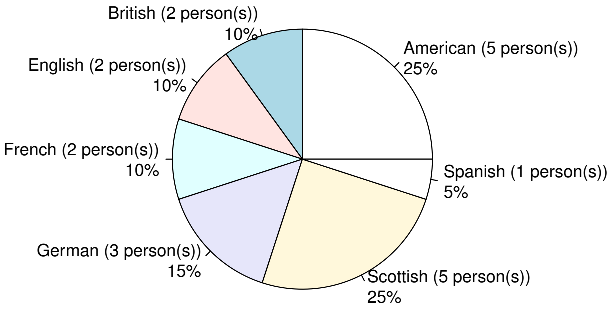


Figure 5: This boxplot depicts the variables f.time, t.time, localisation error and typing error per country.

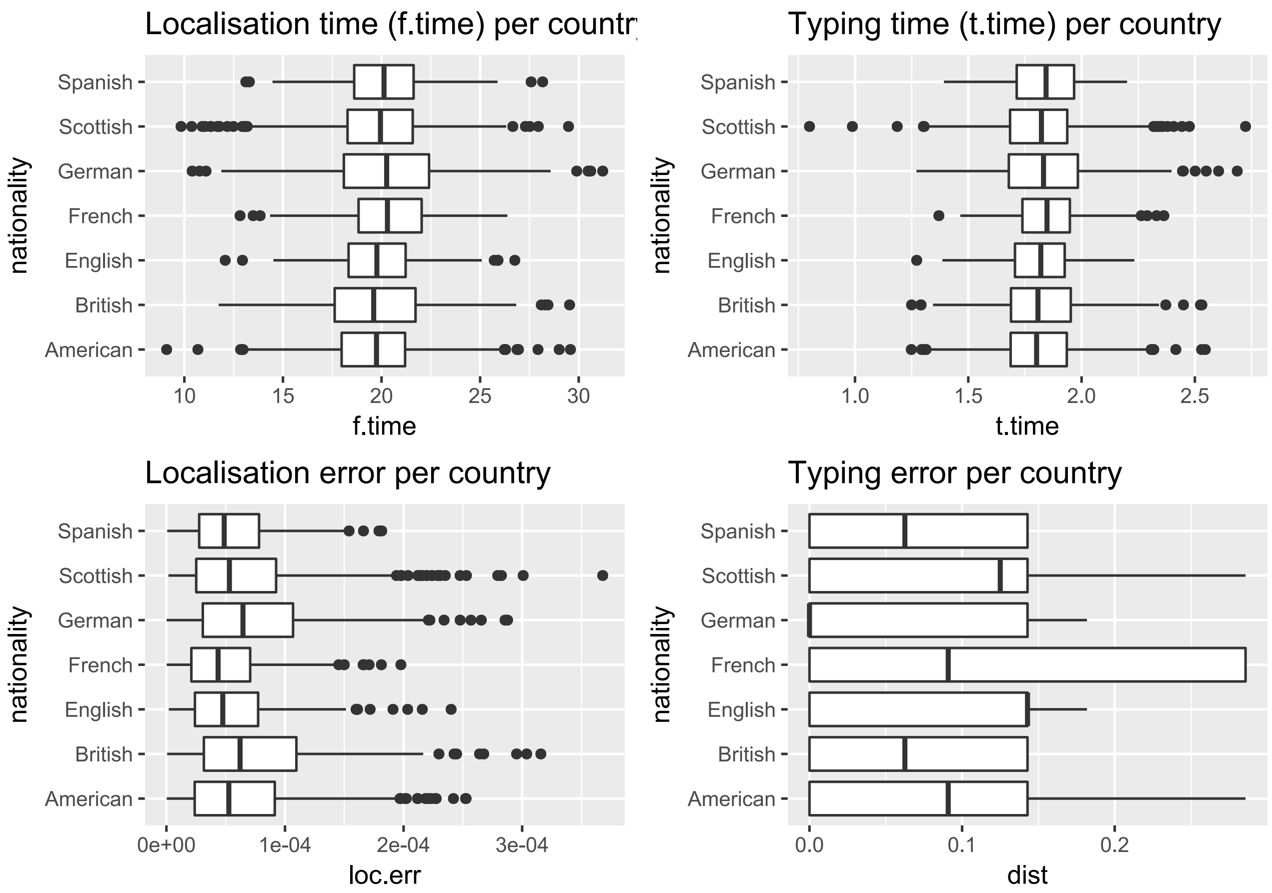


Figure 6: This boxplot depicts the variables f.time, t.time, localisation error and typing error for each gender and in dependency of the used reporting system.

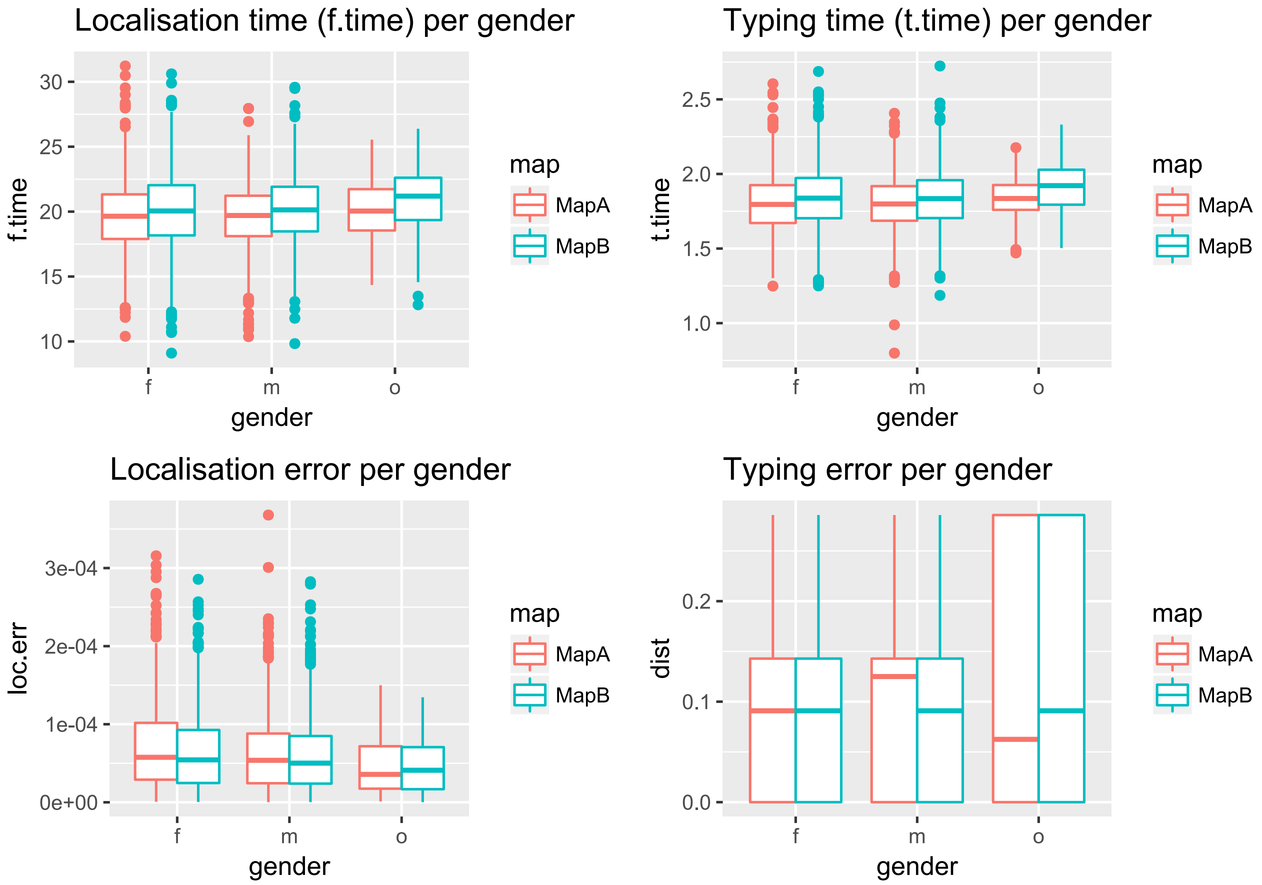


Figure 7: f.time can be regarded as normally distributed. The red line shows the density function of the distribution while the blue, dashed line depicts the reference normal distribution. The histogram shows the relative frequency distribution for a windows size of 0.1.

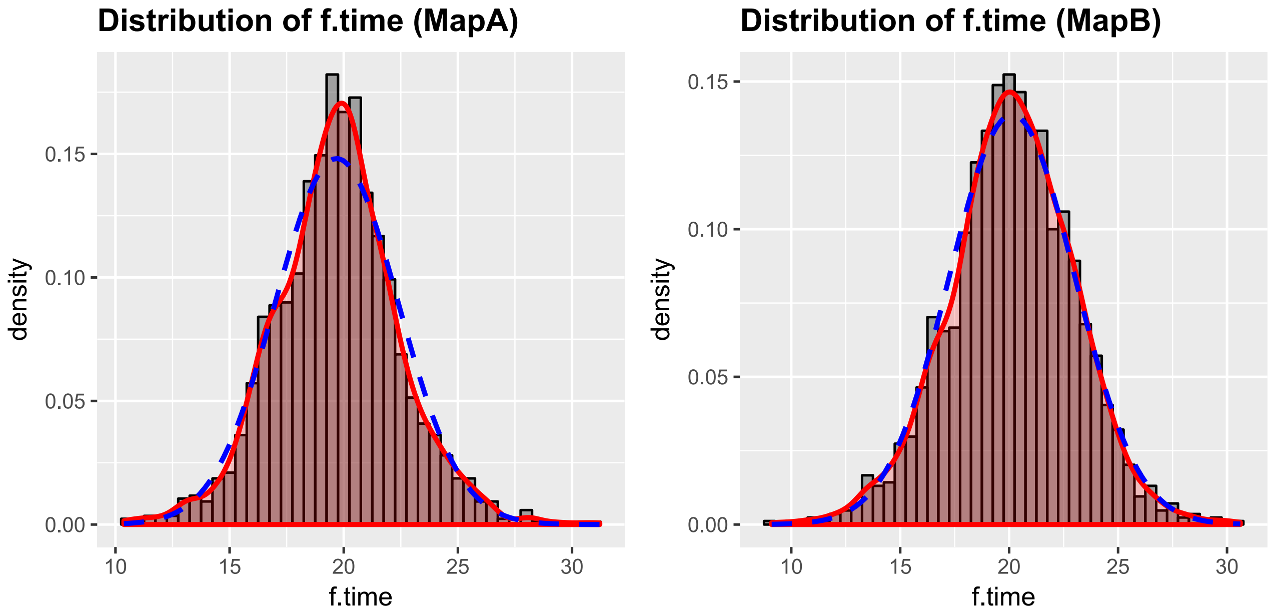


Figure 8: t.time can be regarded as normally distributed. The red line shows the density function of the distribution while the blue, dashed line depicts the reference normal distribution. The histogram shows the relative frequency distribution for a windows size of 0.1.

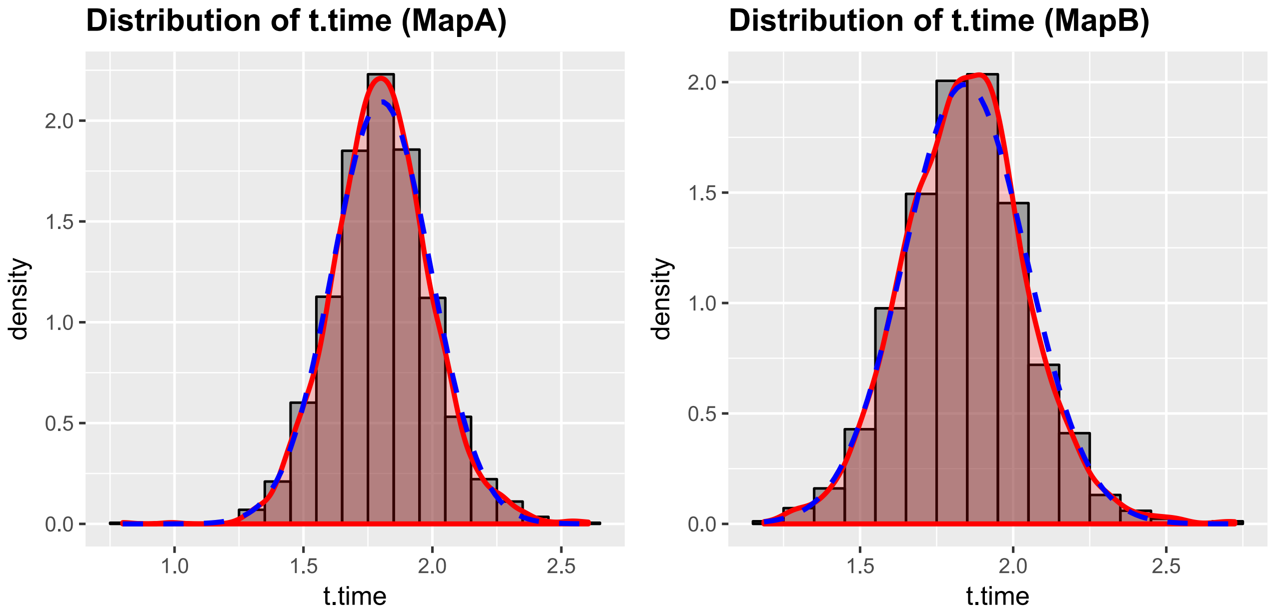


Figure 9: The location error cannot be regarded as normally distributed for none of the two systems.

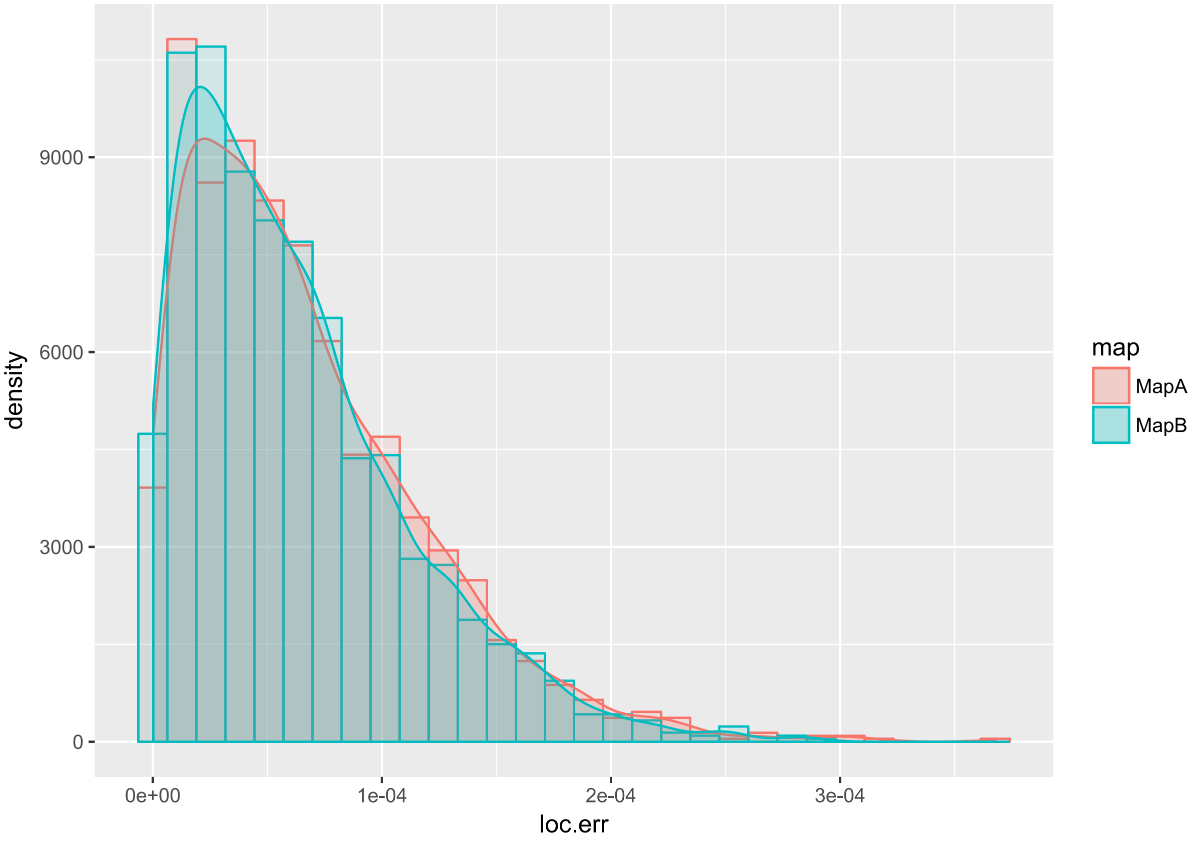


Figure 10: The typing error cannot be regarded as normally distributed for none of the two systems.

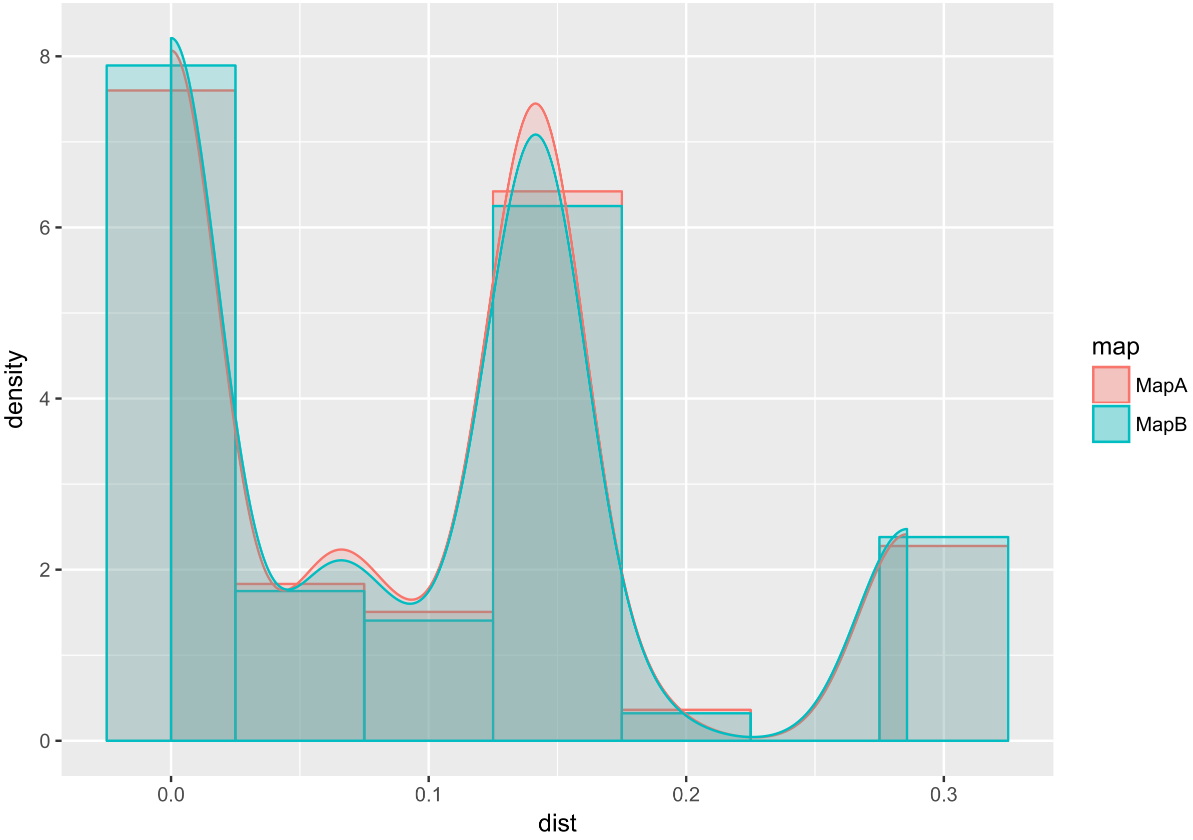


Figure 11: The correlations between each time and error are depicted in the four plots. The blue regression lines show the trends. The R values are discussed in section 5.5.

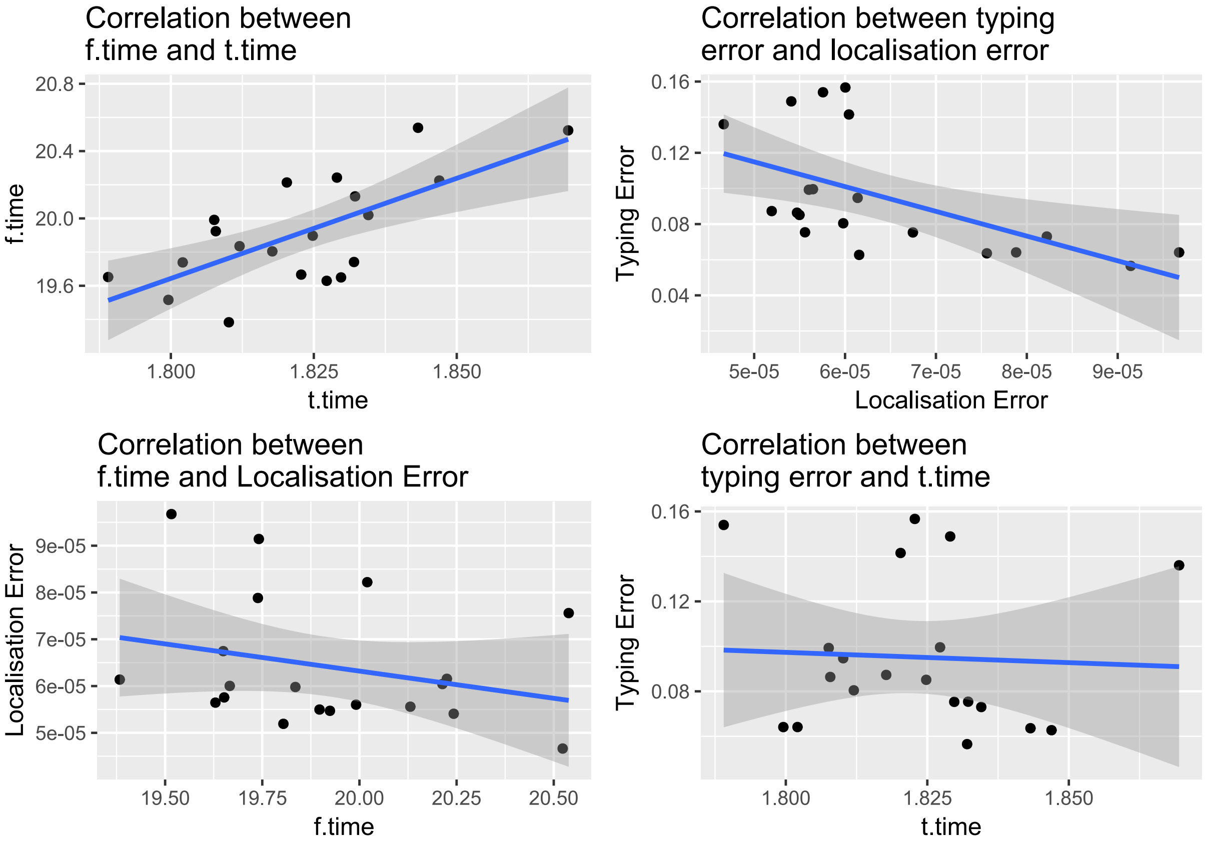


Figure 12: While the age of the participants do not correlate with the times, they strongly correlate with the errors. This is discussed in section 5.5.

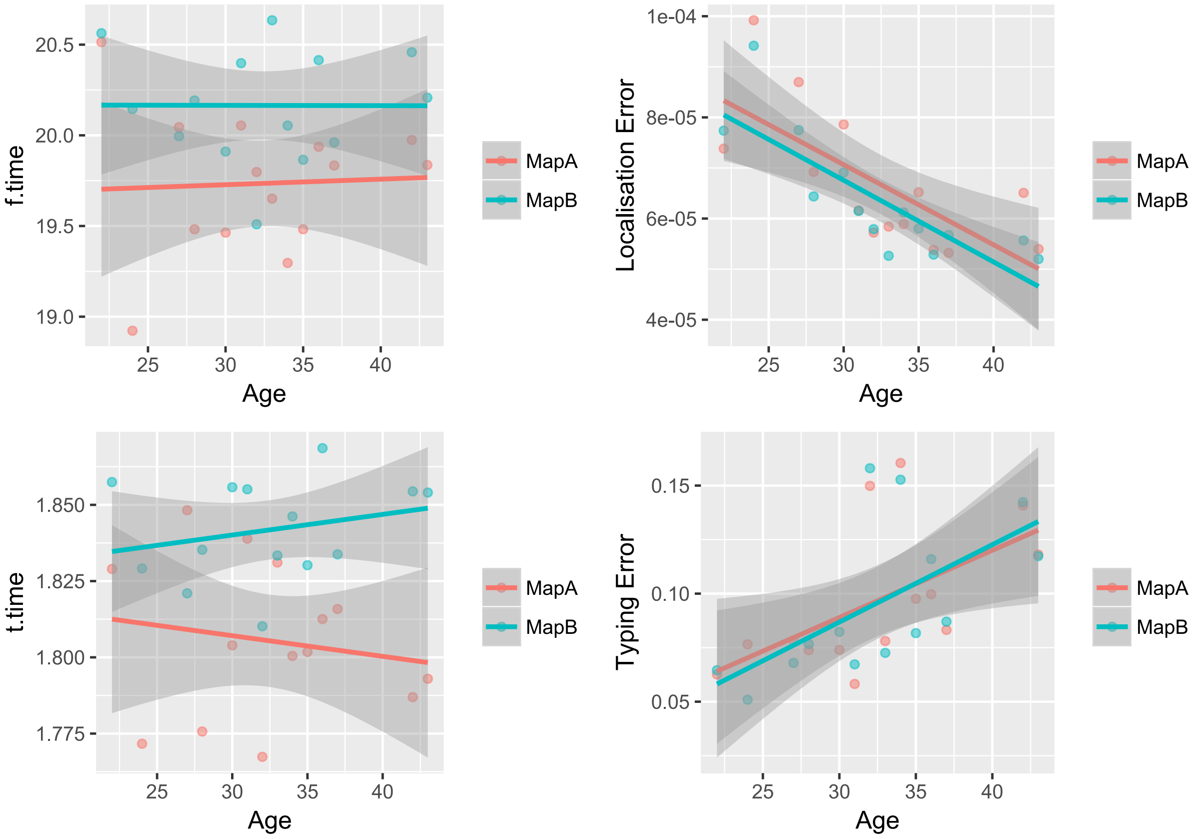
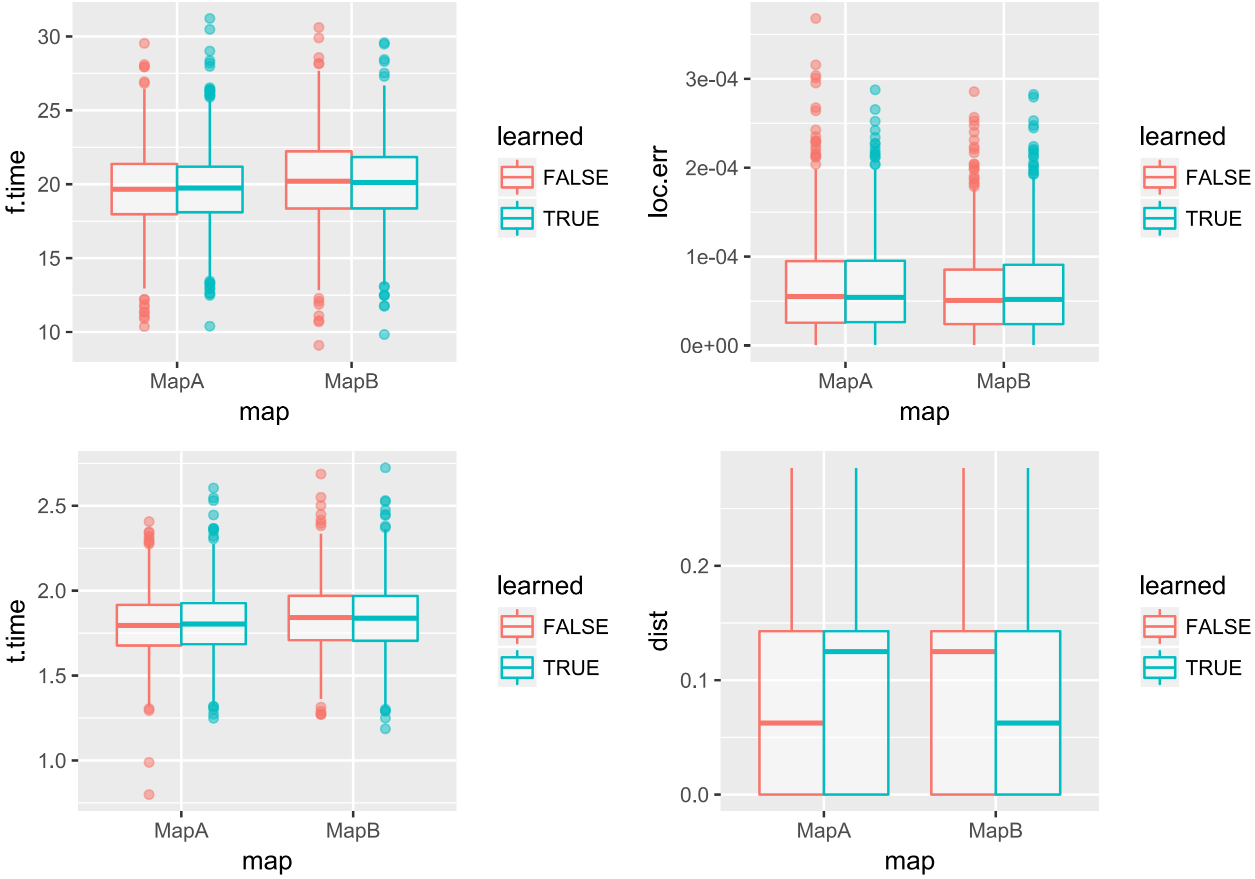


Figure 13: A significant difference between typing error could be found for MapB, which is discussed in section 5.6. Participants who tested MapA first, had a smaller typing error for MapB than people who tested MapB first.


f.time
MapA
MapB
map
learned
FALSE
TRUE
loc.err
3e-04
2e-04
1e-04
0e+00
MapA
MapB
map
learned
FALSE
TRUE
t.time
MapA
MapB
map
learned
FALSE
TRUE
dist
MapA
MapB
map
learned
FALSE
TRUE