A Pragmatic Introduction to Some Common Analyses

CSE 510 – Advanced Topics in Human-Computer Interaction

See due dates and submission information on the course website

Description

You will gain basic familiarity with analyzing experiments using mixed-model analyses of variance in the R statistical package. Consistent with the perspective shared in lecture, this assignment is not intended to provide complete knowledge of how to design or analyze experiments. That goal is well beyond the scope of any one lecture or assignment. This assignment is instead focused on a pragmatic introduction to analyzing experiments based in designs you might later find useful.

Please consider this assignment in the context of the material covered in lecture, as not all of it is repeated here. Completing the assignment will require R:

https://www.r-project.org/

We highly recommend that you use RStudio, an IDE for analysis in R, available free at:

https://www.rstudio.com/

Compiling a PDF within RStudio using knitr will also require pdflatex. RStudio suggests:

On Windows: MiKTeX (Complete):
- http://miktex.org/2.9/setup
- RStudio warns to download the Complete rather than Basic installation, which is available within the "Net Installer". The "Basic Installer" seems to work for James, and is much faster to install.

On Mac OS X: TexLive 2013 (Full)
- http://tug.org/mactex/
- RStudio warns that a download with Safari rather than Chrome is strongly recommended.

On Linux: Use system package manager

Additional Optional Resource: ps4hci

You might also benefit from working through portions of *Practical Statistics for Human-Computer Interaction*, an independent study created by Jacob Wobbrock:

https://depts.washington.edu/madlab/proj/ps4hci/

The first three sections provide an introduction to basic statistical concepts, how to interpret data, and analyses of variance. These sections require you to do some independent research in order to complete the questions (e.g., online, using a statistics textbook). The fourth section is structured more as a tutorial and gives a solid introduction to several types of analyses, including mixed-model analyses used in this assignment. Depending on your existing knowledge, you may be able to skip some or all of Sections 1 to 3 to focus only on the mixed-model portions of Section 4. An answer key is provided, and you are not required to submit any work from this guide. Note the independent study is not in R, but the concepts generalize.

Working with a Partner

When analyzing data (e.g., to write a paper), it is often valuable to talk through your analyses with another person. This is useful for checking that what you did sounds correct and for thinking about how to proceed if stuck. You are welcome to work with a partner throughout this assignment.

Only one person in the partnership needs to submit an assignment. If you work with a partner, please include their name near the top of your report. Please indicate in your submission if you talked through the analysis with others who were not your partner.

Data Files and Formatting

You will work with three datasets: one artificial and two from actual published studies. The data is appropriate for these analyses, but explicitly not cleaned up for the sake of this assignment. One important implication is that you need to be mindful of the types assigned to columns in provided data files. It will be your responsibility to decide whether each field should be *continuous* (i.e., *numeric* in R) or *nominal* (i.e., *factor* in R). You may safely choose to ignore *ordinal* typing for the purposes of this assignment, but may want to learn about it at some point in the future.

You will use R Markdown to complete the assignment. R Markdown will create a PDF that contains your comments, scripting commands, and output. This is a good practice for documenting your analysis, and easier than trying to copy and paste commands and graphs into a different document. It also allows clean interleaving of written components. To create a new R Markdown template file with instructions in R Studio: *File→New File→R Markdown*. However, for this assignment, you will want to work with our provided starter in *cse510statslab.rmd*.

Coordination and Submission Procedure

You should submit a report in PDF addressing the bulleted questions posed in the body of this assignment (items 1.1 and 1.2, 2.1 through 2.10, 3.1 and 3.2), together with the source files you used to generate the PDF (e.g., your R Markdown files). We have provided *cse510statslab.rmd* to scaffold this PDF generation. Submit a ZIP containing your PDF and source file via Canvas.

Grading

You will be graded on the *correctness* and the *appropriateness* of your responses to questions posed by the assignment. The notion of *correctness* is hopefully self-evident, though we have noted that community norms can sometimes make this less obvious and we will work to give reasonable credit for reasonable responses. Regarding *appropriateness*, grading will be based in striking an appropriate balance between reporting *sufficient* detail and reporting *excessive* detail. The goal is to gain experience reporting your results in approximately the same level of detail that should be included in reporting a research result.

Parametric vs Nonparametric Tests

This lab suggests using parametric tests to analyze Likert data, which some people consider inappropriate because Likert data may not satisfy data distribution assumptions. Other work suggests parametric tests can be used without fear of the "wrong" answer, even for Likert data[1].

[1] Geoff Norman. Likert Scales, Levels of Measurement and the "Laws" of Statistics. Advances in Health Sciences Education 15, 5. 2010. https://doi.org/10.1007/s10459-010-9222-y

We will accept either approach, as long as that approach is appropriately applied and rationalized, but the lab teaches and assumes the use of parametric tests for all examined data.

Study 1: Text Input Method Words Per Minute

This section will walk through an example analysis to introduce you to the concepts. It uses a small fictitious dataset from a previous version of Wobbrock's guide. It represents a within-subjects experiment measuring text input speed of ten participants using three interfaces (speech, keyboard, Graffiti). Wobbrock gives the following backstory for this fictitious data:

> *The study compared three text input methods: Graffiti, keyboard typing, and speech recognition. After fifteen minutes of practice with each method, participants entered 20 phrases with each method. For each participant, trials with WPM less or more than 2 standard deviations from the participant's mean of trials were removed as outliers. The measure for each technique was the average of the non-outlier phrases for each method for each participant.*

Read the Data

First, be sure to run the **`setup`** code chunk at the top of the file. If you do not have the required packages installed, uncomment the "install.packages" command.

Next open the dataset in R and inspect it to see what type each variable is. This code is included in code chunk **`study1_readData`** in the markdown file.

1.1 What type was each variable assigned to? Are these "correct"?

We can convert data types to reassign them appropriately using `as.factor()` to convert to a nominal type or `as.numeric()` to convert to a continuous type.

Run code chunk **`study1_fixTypes`** to correct the types.

Examine *WPM*

Now we will examine the variable *WPM*, both via descriptive statistics and plots. We do this both to get familiar with the dataset and to get familiar with capturing screenshots of our exploration and analysis of datasets in R. There are multiple plotting libraries, but this example will use qplot in the ggplot2 library. Run code chunk **`study1_examineWMP`** to examine the outputs.

Plot *WPM* vs *Participant*

Although the plot shows the different values of *WPM*, it does not include any information about the relationship of *WPM* to the different interfaces.

If *method* were the only thing that had impacted *WPM*, we might base our analysis by simply splitting *WMP* out by interface. But there is also a potential for *participant* to have impacted *WPM*, as some people might be slower or faster (regardless of which interface they are using). When we plot the values of *WPM* by *participant*, as in code chunk **`study1_WMPVParticipant`**, we see this does seem to be true (with participants means ranging from as low as 65 to as high as 102).

Create and Test a Mixed Model

Given the impact of *participant* on *WMP*, our challenge is to analyze the data in a way that allow us to determine whether different *methods* actually enable faster input (or if the difference is instead simply due to individual variations).

In the language of mixed-model analyses, *method* is a fixed effect. It has three levels that were selected because they are of interest (we are interested in the effect of speech, keyboard, and Graffiti). In contrast, *participant* is a random effect. Participants were randomly sampled from a larger population over which we wish to generalize (we want to know whether *methods* are faster for the larger population) but we are not interested in whether P1 or P2 was faster.

We therefore analyze our data with the `lmerTest` package, which runs *linear-mixed effect models*. It's based off the `lme4` package, but includes additional tests for significance in the output. See code chunk **`study1_mixedModel`** to create and test the model.

A few things to notice. `lmer` fits a linear, mixed effect model. There are other packages for running different types of models, such as `lm` for a linear model with only fixed effects (`stats` package) and `clmm` for fitting cumulative link mixed models, which are more appropriate for ordinal responses (`ordinal` package).

Within `lmer`, first the response variable is identified (`WPM`), followed by a tilde (`~`). Effects are then separated by `+` by their column name. Note `Participant` is specified as a random effect with the notation `(1 | Participant)`.

We can then summarize the model and run an `anova` on the fixed effects.

The first thing we will look at here is the anova results. Here we can see we only have one variable, *Method*. This test shows us that *method* has a statistically significant impact on *WPM*. We can also see the least squares mean values for the different levels of *method* in the "Fixed Effects" section of the model summary, but there has not yet been any test of whether these differences are significant. We will treat this as an unplanned comparison.

Run a Tukey HSD

An appropriate test to run to examine the differences between the methods would be a Tukey HSD, which finds significant differences between means. Run code chunk **`study1_tukey`** to do so.

`glht` stands for General Linear Hypotheses, which is useful for making multiple comparisons on models including linear mixed-effect models. `linfct` is used to specify the hypothesis to test, which takes a multiple comparisons, or `mcp` object. Note `Method` in `mcp` refers to the variable/column name.

We can see by the fact that the differences are all significant that speech, keyboard, and Graffiti are all significantly different from each other (all three pairs of tests are significant).

Describe Your Analysis

1.2 Complete the description of your analysis in your R Markdown file.

We performed a mixed-model analysis of variance, treating *method* as a fixed effect and *participant* as a random effect. We found a significant main effect of *method* ($F($ ____ , ____$) =$ ____, $p \approx$ ____), prompting us to investigate pairwise differences. We employed Tukey's HSD procedure to address the increased risk of Type I error due to unplanned comparisons, finding that *speech* leads to significantly greater *WPM* than both *keyboard* ($z =$ ____, $p \approx$ ____) and *Graffiti* ($z =$ ____, $p \approx$ ____) and that *keyboard* also leads to significantly greater *WPM* than *Graffiti* ($z =$ ____, $p \approx$ ____).

Study 2: Comparison of Multiple Interfaces

This data was published in the following UIST 2007 paper:

> Raphael Hoffmann, James Fogarty, Daniel S. Weld. (2007). Assieme: Finding and Leveraging Implicit References in a Web Search Interface for Programmers. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2007)*. pp. 13-22.

This is a real dataset collected in a within-subjects experiment comparing participant completion of ten tasks in each of three interfaces. The completion of a task in any interface rendered it useless for testing the other interfaces (i.e., knowing the answer from completing a task once rendered that task a poor measure of the other interfaces). The experimenters therefore assembled a library of forty tasks, presented ten random tasks to participants during an initial practice stage, and then presented ten random tasks for each of the three interfaces. Each participant therefore completed each of the forty tasks exactly once. Interfaces were presented using a counterbalanced design.

The data file contains eleven columns:

Independent Variables

Participant: A unique identifier for the participant

Trial: Which trial this is for the participant (range 1 to 30)

Interface: Which interface the participant is using (A, B, C)

Task: A unique identifier for the task

TaskSel: Whether or not the task has the Sel property, an important type of task (0 or 1)

TaskCoo: Whether or not the task has the Coo property, an important type of task (0 or 1)

TaskUse: Whether or not the task has the Use property, an important type of task (0 or 1)

TaskEx: Whether or not the task has the Ex property, an important type of task (0 or 1)

Dependent Variables

Restarts: A count of the number times a participant chose to restart the task

Time: Total time spent on the task

Correctness: An expert rating of the quality of the participant's solution to the task

You will find the instructions provided here are much less detailed than those for Study 1. This is intentional, with the instructions intended only as high-level guidance through this analysis.

Read the File

2.1 What type was assigned to each variable? Did you need to change any types?

Analyze Time

Run a mixed-model analysis of variance for *Time*, as estimated by the independent variables.

2.2 Create a mixed-effect model for *Time* and run an ANOVA

You will notice that a number of the independent variables above have no significant impact on *Time*. It is common to therefore remove them from the remainder of your analysis of *Time*.

Document this and run an new mixed-model analysis of variance for *Time*, using only the variables that were significant.

2.3 Create a mixed-effect model for *Time*, using only the significant variables, and run an ANOVA

Trial has a significant effect (if not, check your work up to this point), but is not our variable of interest. The negative parameter estimate means that people took less *Time* to complete the task as *Trial* increased (maybe they got better at the task, maybe they got sick of it and did not try as hard). What we should be concerned about is whether *Trial* affects *Interface*. Add an additional parameter to the model, created by crossing *Trial* with *Interface*, (i.e., adding Trial*Interface to the model). This is an interaction.

2.4 Create a model including *Trial* crossed with *Interface* and run an ANOVA

2.5 Does *Trial* significantly interact with *Interface*? What does this mean?

Go back to working with the same variables you had in 2.3. *Interface* has a significant effect, so test significance of the difference between the levels of *Interface* and obtain the pairwise contrasts.

2.6 Complete your analysis of *Time* and *Interface*.

Once again, you have now captured everything you need to report an analysis of *Time*.

2.7 Prepare a description of your analysis that resembles that in 1.6.

Note that your analysis here is more complex and the resulting description will be longer. After all, you removed variables that were not significant, checked an interaction, interpreted that interaction, and only then conducted your analysis of the *Interface* variable.

Analyze Restarts

Run a mixed-model analysis of variance for *Restarts*, as estimated by the independent variables.

2.8 Prepare a description of your analysis of Restarts that resembles that in 1.6.

Here you are asked to perform the entire analysis on your own.

Analyze Correctness

Run a mixed-model analysis of variance for *Correctness*, as estimated by the independent variables.

2.9 Prepare a description of your analysis of Correctness that resembles that in 1.6.

Again you are asked to perform the entire analysis on your own.

Summarize the Results

You have now analyzed three different independent measures each intended to give some insight into the appropriateness of the three different interfaces studied here. Summarize your results and indicate which interface seems to be the best for these tasks.

2.10 Summarize your overall results.

Study 3: Comparison of Multiple Conditions

This data was published in this CHI 2016 paper:

> Xiaoyi Zhang, Laura R. Pina, James Fogarty. (2016). Examining Unlock Journaling with Diaries and Reminders for In Situ Self-Report in Health and Wellness. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2016).*

This is a real dataset collected in a within-subjects experiment comparing participant data logging with six conditions (two interfaces, three notification conditions). Participants selected the type of data they were most interested in journaling (sleepiness, pleasure and accomplishment, or mood). Participants used an Android application over 18 days. Each condition was used for two consecutive days, with a rest day between conditions. Conditions were presented using a counterbalanced design.

The two interfaces consisted of an app which participants could open and log data from (*without unlock*), and a lock-screen interaction for logging data (*with unlock*). Participants either received no notifications (*none*), *traditional* notifications consisting of a push notification, or *aggressive* notifications which added sound and vibration to the push notification. Participants could journal at any time from within the app in all conditions. This was the only option in the *none, without unlock* condition. Participants were asked to journal every 30 minutes for a 12-hour window. Notifications were delivered every 30 minutes in the two notification conditions.

The data file contains eleven columns:

Independent Variables

Participant: A unique identifier for the participant

Gender: The gender the participant identified as

Type: The type of data the participant entered (S = Sleepiness, P = Pleasure, M = Mood)

CalendarDay: Calendar day since the beginning of the study

StudyDay: Day since the beginning of the study only considering logging days

WithUnlock: Whether or not the participant had the lock-screen interaction that day (0 or 1)

Notification: Notifications received that day (N = None, T = Traditional, A = Aggressive)

Dependent Variables

Intrusiveness: Likert rating for interaction intrusiveness that day (1 to 5)

Frequency: Number of times participant logged data that day

Timeliness: Average number of minutes to nearest journaling interval (-15 to 15)

These instructions are even less detailed than the previous two studies, but more closely approximate a real-world analysis.

Analyze the Data

Analyze the data and find what contributes to the dependent variables. It is up to you to determine the types of independent variables and select an appropriate model to fit. You must show us that you analyzed this data, but do not need to describe your analysis in detail.

3.1 Analyze the data.

Summarize the Results

You have now analyzed three different independent measures which provide some insight into the six study conditions. Summarize your results as you might see in a published paper. Of course, you could read simplify the final paper and see how the authors summarized their analysis. But treat this as an exercise in learning to describe the results yourself.

3.2 Summarize your overall results.