A Pragmatic Introduction to Some Common Analyses

CSE 510 – Advanced Topics in Human-Computer Interaction

DUE: Sunday, May 14

Description

You will gain basic familiarity with analyzing experiments using mixed-model analyses of variance in either the JMP or R statistical packages. Consistent with lecture, this assignment is not intended to provide complete knowledge of how to design or analyze experiments (a topic far beyond the scope of 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.

In addition to my lecture material and the contents of this assignment, you might benefit from working through the first four sections of *Practical Statistics for Human-Computer Interaction*, an independent study created by Jacob Wobbrock and linked from the course webpage:

https://depts.washington.edu/aimgroup/proj/ps4hci/ps4hci.zip

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., 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-models portions of Section 4. An answer key is provided, and you are not required to hand in any work from this guide.

This assignment was originally developed using the JMP statistical package. R has since also become popular and mature. We highly recommend that you use these with an IDE for analysis. Images in the JMP version of this assignment are from Version 7, but the functionality is the same. You are also free to use any other package, but you will likely find this assignment much more difficult to complete and we will not be able to provide any assistance in completing it correctly.

JMP includes an IDE, and is available as a free trial, or through a UW CSE discount:

http://www.jmp.com/en_us/software/jmp.html

https://e5.onthehub.com/WebStore/AdTargetOfferingList.aspx?wsmv=7ed98060-98ea-e411-940b-b8ca3a5db7a1&ws=a4fce2bc-ac2d-de11-a497-0030485a8df0&vsro=8&utm_source=jmp-version12-rs&utm_medium=WebStoreAds&utm_campaign=JMP

RStudio is a free IDE available at:

https://www.rstudio.com/

Working it Through 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. The primary implication is that you need to be mindful of the types which your package assigns to columns when you load data files. It will be your responsibility to decide whether each field should be *continuous* or *nominal* (you can safely ignore the *ordinal* type for this assignment).

If using JMP, note you should likely save the provided CSV files into the JMP format, or your type information will be lost when you later come back to the file.

If using R, note R is an interpreted language. As a result, people tend to interleave scripting in a file and running commands on the command line. We have provided a sample script file to demonstrate the analyses in the artificially-generated dataset.

If using R, the easiest formatting solution is to use R Markdown or compile your R notebooks. This will create a PDF that contains your comments, scripting commands, and output rather than trying to copy and paste graphs, commands, etc. into a different document. R markup language will allow for the cleanest interleaving of the written components as well.

To compile your notebook in R Studio: *File→Compile Notebook*

To get started with R Markdown in R Studio: *File→New File→R Markdown…* will create a template with instructions to get you started.

Coordination and Submission Procedure

You should submit a report in PDF or HTML addressing the bulleted questions posed in the body of the assignment (items 1.1 through 1.6, 2.1 through 2.13, 3.1 to 3.2). Duplicate the bulleted question in your report, but not the explanatory text between questions. The assignment is available in both Word and PDF formats so that it is easier for you to duplicate the questions.

Be aware of the need to preserve high-resolution images in your electronic submission. If you submit a PDF including screenshots, ensure resolution is preserved so that we can zoom in to the point of being able to read any details. This is most likely to be a concern if you are outputting to PDF with settings that compress images.

Submit a PDF or a ZIP of your work 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. 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.

Study 1: Text Input Method Words Per Minute

This is 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 different interfaces (speech, keyboard, and Graffiti). Wobbrock gives the following backstory for understanding this fictitious data:

> *The study compared three text input methods, Graffiti, keyboard typing, and speech recognition. After fifteen minutes of practice with each, subjects entered 20 phrases with each method. For each subject, trials whose WPM were less or more than 2 standard deviations from the subject'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.*

Open the dataset in R and inspect it to see what type each variable is.

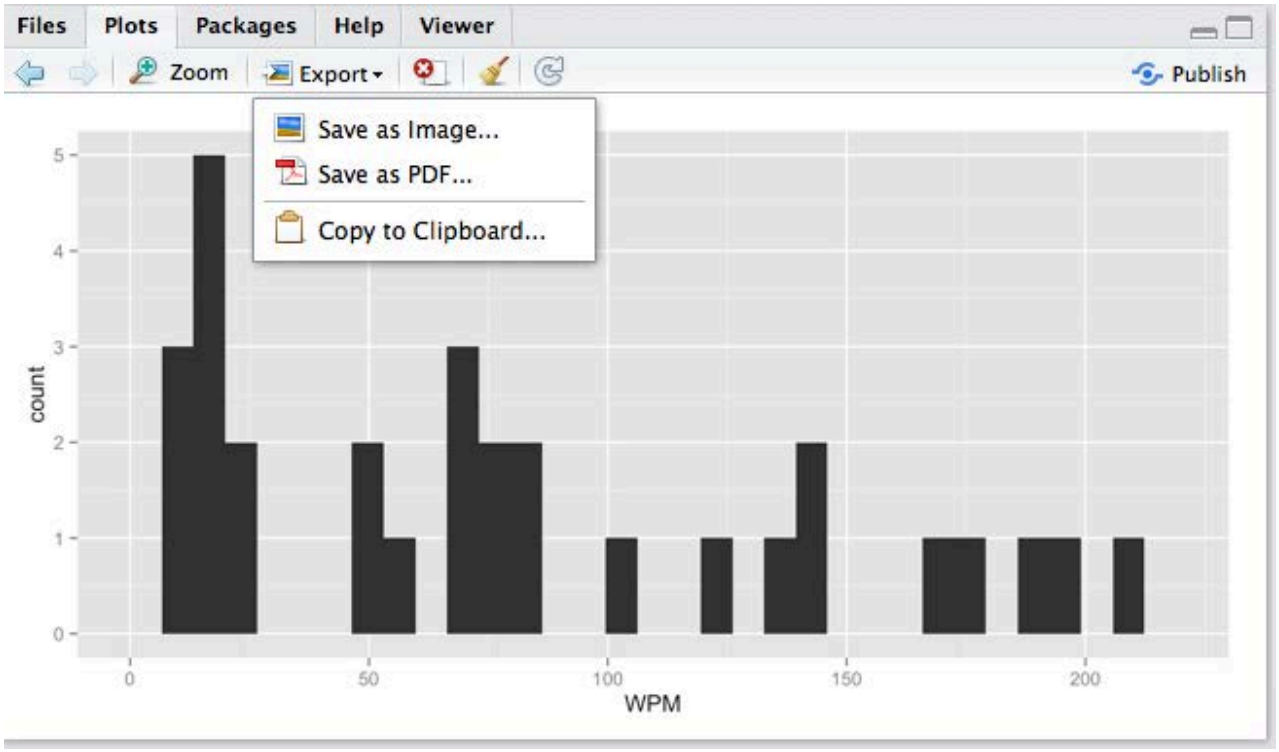
```
> study1.data <- read.csv("Study1.csv")
> str(study1.data) # See what type each variable is.
'data.frame':30 obs. of  3 variables:
 $ Participant: int  1 1 1 2 2 2 3 3 3 4 ...
 $ Method     : Factor w/ 3 levels "Graffiti","keyboard",..: 1 2 3 1 2 3 1 2 3 1 ...
 $ WPM        : num  13.2 55.7 143.3 14.2 68.3 ...
```

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

We can convert data types using `as.factor()` or `as.numeric()` to reassign appropriately (see the script for an example).

Now we will plot *WPM*. 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.

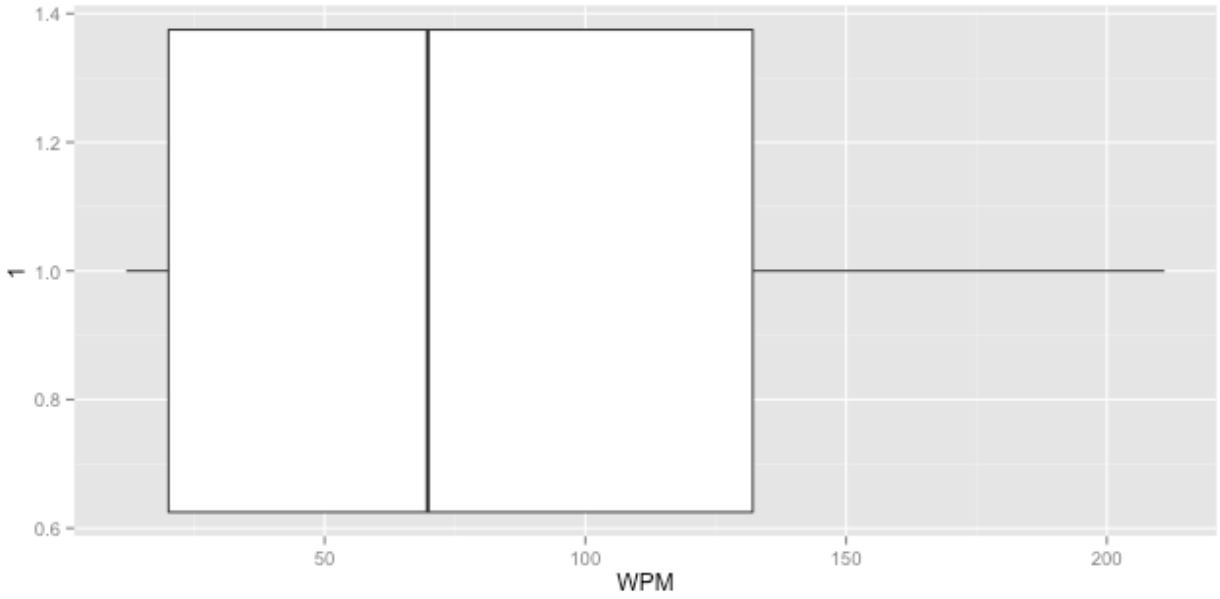
In R, this is a few lines of code (see below). There are multiple plotting libraries, this example will go through using qplot in the ggplot2 library. Exporting plots generated can be done either using the ggsave command or by by clicking "Export" under the "Plots" menu.



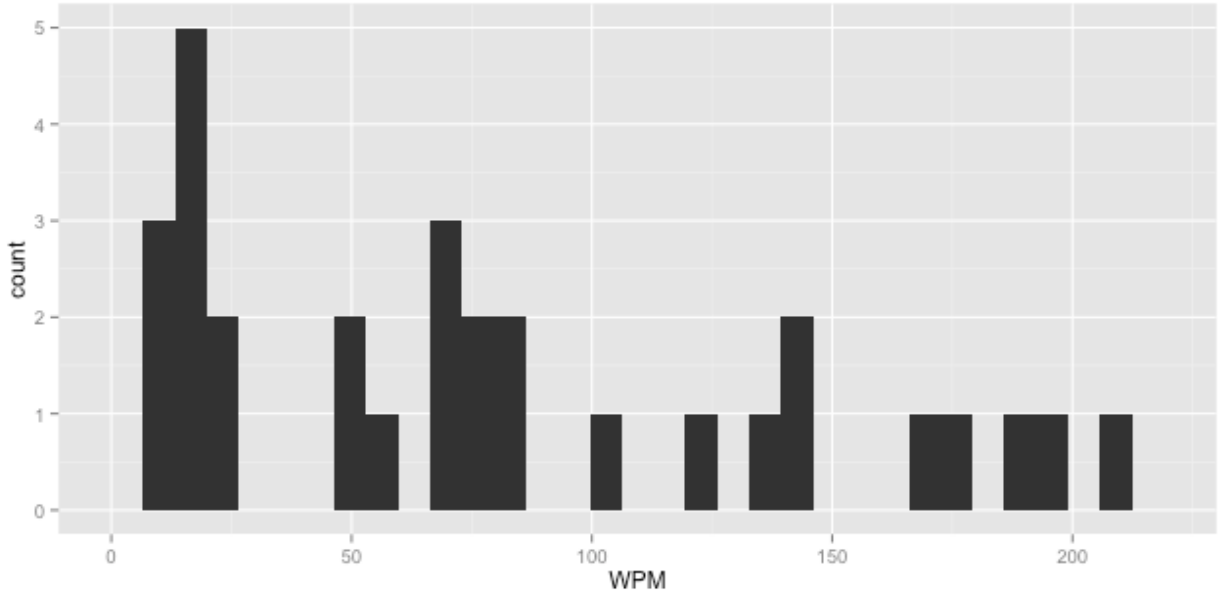
1.2 Insert the commands you used to plot and summarize the values of WPM.

In R, you should get results that looks like this when you execute the following commands.

```
> summary(study1.data$WPM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.97   20.07   69.78   80.97  132.10  211.00
> library(psych)
> describe(study1.data$WPM)
  vars  n  mean   sd median trimmed   mad   min    max  range skew kurtosis    se
1    1 30 80.97 63.2  69.78    74.8 77.27 11.97 211.05 199.08 0.58    -1.01 11.54
> library(ggplot2)
> qplot(x=1, y=WPM, data=study1.data, geom="boxplot") + coord_flip()
```




```
> qplot(WPM, data=study1.data, geom="histogram")
```



Make sure you're comfortable copying/formatting code and images.

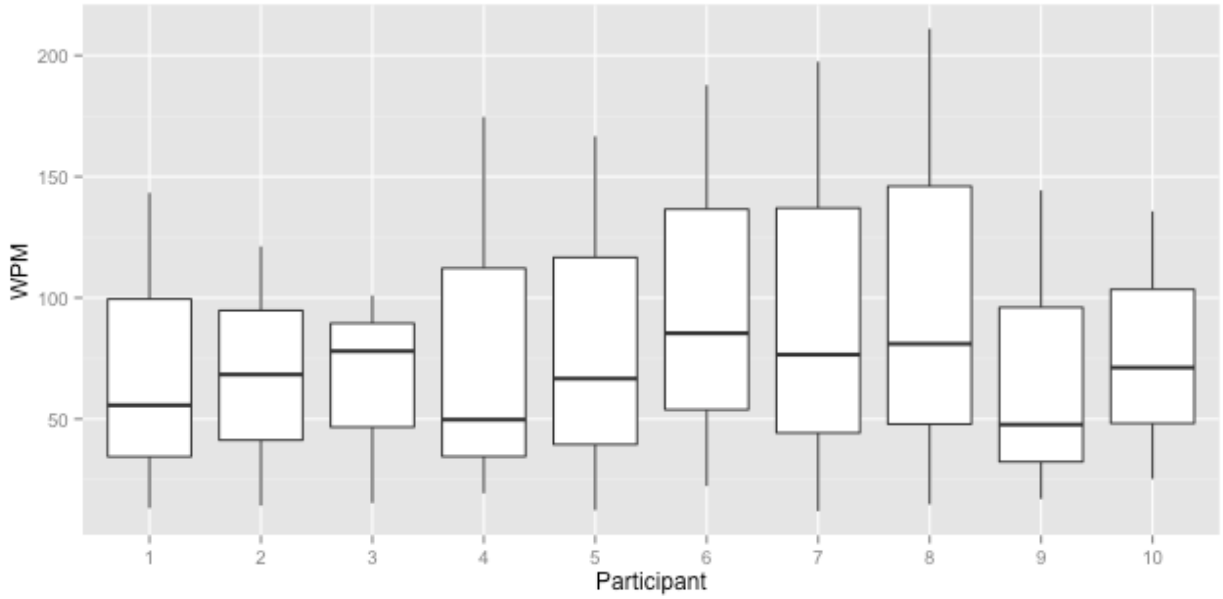
1.3 Insert a screenshot of plot of the values of WPM.

That plot does not seem to tell us much about the relationship of *WPM* to the different interfaces. So now we will split out *WPM* by interface. We can add a second variable for *method* to our analysis, see the examples below.

If *method* were the only thing that had impacted *WPM*, we might base our analysis on such a division of the data. 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). Plot the values of

WPM by *participant*, and you can see that this does seem to be true (with participants having means ranging from as low as 65 to as high as 102). So our challenge is to analyze this data in a way that allows us to determine whether different *methods* actually enable faster input (or if the difference is instead simply due to individual variations).

```
> qplot(x=Participant, y=WPM, data=study1.data, geom="boxplot")
```



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

```
> library(lmerTest)
```

```
> study1.mixedmodel <- lmer(WPM ~ Method + (1 | Participant), data=study1.data)
```

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)`.

1.4 Insert the commands you used to create a mixed-effect model.

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

```
> study1.mixedmodel
Linear mixed model fit by REML ['merModLmerTest']
Formula: WPM ~ Method + (1 | Participant)
   Data: study1.data
REML criterion at convergence: 249.8256
Random effects:
 Groups      Name        Std.Dev.
 Participant (Intercept)  6.035
 Residual                 20.958
Number of obs: 30, groups:  Participant, 10
Fixed Effects:
   (Intercept)  Methodkeyboard    Methodspeech
         16.57           51.46          141.76
> anova(study1.mixedmodel)
Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
       Sum Sq Mean Sq NumDF  DenDF F.value    Pr(>F)
Method 102989   51494     2 18.001  117.24 4.756e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The first thing we will look at here is the anova results. Here we can see we only have one variable, *Method*. If we had accidentally forgotten to denote *participant* as a random effect, it would also appear here. 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 both *method* and *participant* in the “Fixed Effects” section, but there has not yet been any test of whether these differences are significant. We will treat this as an unplanned comparison. An appropriate test to run would be a Tukey HSD.

Running one gives us:

```
> library(multcomp)
> study1.tukey <- glht(study1.mixedmodel, linfct=mcp(Method = "Tukey"))
> summary(study1.tukey)


       Simultaneous Tests for General Linear Hypotheses


Multiple Comparisons of Means: Tukey Contrasts
```

```
Fit: lme4::lmer(formula = WPM ~ Method + (1 | Participant), data = study1.data)


Linear Hypotheses:
                          Estimate Std. Error z value Pr(>|z|)
keyboard - Graffiti == 0   51.455      9.373   5.490 1.07e-07 ***
speech - Graffiti == 0    141.756      9.373  15.124  < 1e-07 ***
speech - keyboard == 0     90.301      9.373   9.634  < 1e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

`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).

1.5 Insert the output of your final model for analyzing method and WPM.

This analysis was very straightforward, and all of the information you need to report it is included in your screenshots 1.4 and 1.5.

1.6 Close R and use your 1.4 and 1.5 text to complete this paragraph, which assumes the actual means are presented somewhere else in the paper (such as a table).

We performed a mixed-model analysis of variance, treating *method* as a fixed effect and *participant* as a random effect. The omnibus test showed a significant main effect of *method* ($F($ ____ , ____ $) =$ ____, $p <$ ____), 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* ($z =$ ____, $p <$ ____) and *Graffiti* ($z =$ ____, $p <$ ____) and that *keyboard* also leads to significantly greater *WPM* than *Graffiti* ($z =$ ____, $p <$ ____).

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 (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, but hopefully still contain enough to guide you on a correct path.

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

Analyzing Time

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

2.2 Insert the commands you used to create a mixed-effect model.

2.3 Insert a screenshot of the resulting output.

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*. Run a new mixed-model analysis of variance for *Time*, using only the variables that were significant.

2.4 Insert the commands you used to create this model.

2.5 Insert a screenshot of the resulting output.

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*. This is an interaction.

2.6 Insert the commands you used to create this model.

2.7 Insert a screenshot of the resulting output.

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

Go back to working with the same variables you had in 2.4. *Interface* has a significant effect, so test significance of the difference between the levels of *Interface* and obtain the pairwise contrasts.

2.9 Insert a screenshot of your final dialog analyzing *Time*.

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

2.10 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.

Analyzing Restarts

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

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

Here you are asked to perform the entire analysis on your own. Only the final description is strictly necessary, but it will be easier to award partial credit if you show your work.

Analyzing Correctness

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

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

Again you are asked to perform the entire analysis on your own. Again only the final description is strictly necessary, but it will be easier to award partial credit if you show your work.

Summary

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.

2.13 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 ever 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.

Analysis

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. One way of doing this

would be to attach your R script or notebook for analysis, with a few comments explaining why you did certain things (see the Study 1 R script for an example).

3.1 Show us how you analyzed this data.

Summary

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.