Adaptability Study - Overview of Data Analysis

Kevin A. Ryan

Monday, July 20, 2015

***why reproducible research***

What reproducible reporting is about is a validation of the data analysis. Because you're not collecting independent data using independent methods, it's a little bit more difficult to validate the scientific question itself. But if you can take someone's data and reproduce their findings, then you can, in some sense, validate the data analysis. *This involves having the data and the code because more likely than not, the analysis will have been done on the computer using some sort of programming language, like R.* So you can take their code and their data and reproduce the findings that they come up with. Then you can at least have confidence that the analysis was done appropriately and that the correct methods were used.

* Replication, whereby scientific questions are examined and verified independently by different scientists, is the gold standard for scientific validity.
* Replication can be difficult and often there are no resources to independently replicate a study.
* Reproducibility, whereby data and code are re-analyzed by independent scientists to obtain the same results of the original investigator, is a reasonable minimum standard when replication is not possible.

***The Data Science Pipeline***

* The process of conducting and disseminating research can be depicted as a "data science pipeline"
* Readers and consumers of data science research are typically not privy to the details of the data science pipeline
* One view of reproducibility is that it gives research consumers partial access to the raw pipeline elements.
* \* Scientific Question ->  
  \* Protocol ->  
  \* Experiment ->  
  \* Measured Data -> <- PROCESSING CODE  
  \* Analytic Data -> <- ANALYTIC CODE  
  \* Computational Results -> <- PRESENTATION CODE  
  \* Figures/Tables/Summaries->   
  \* Article

***Elements of Reproducible Research***

[need to refer to a graphic that presents the entrity of the Adaptability work]

1. Analytic data. The data that were used for the analysis that was presented should be available for others to access. This is different from the raw data because very often in a data analysis the raw data are not all used for the analysis, but rather some subset is used. It may be interesting to see the raw data but impractical to actually have it. Analytic data is key to examining the data analysis.
2. Analytic code. The analytic code is the code that was applied to the analytic data to produce the key results. This may be preprocessing code, regression modeling code, or really any other code used to produce the results from the analytic data.
3. Documentation. Documentation of that code and the data is very important.
4. Distribution. Finally, there needs to be some standard means of distribution, so all this data in the code is easily accessible.
5. Literate (Statistical) Programming. Using R and knitR

\*\*\* Steps in Data Analysis \*\*\*

***1. Defining the question***

The purpose of the study is to determine, for the U.S. Army Asymmetric Warfare Group (AWG) quantitative metrics of military-relevant adaptive behavior. We expect a positive relationship between performance on the executive function tasks and military-relevant adaptive test tasks. We also expect a negative relationship between adaptive performance and stress reactivity. Specifically, higher adaptive performance will be negatively correlated with metrics of cardiovascular, neuroendocrine, and electro-dermal stress response indices.

***2. Defining the ideal dataset*** If you want to make inference about a problem then you have to be very careful about the sampling mechanism and the definition of the population that you are sampling from. Typically, when you make an inferential statement, you use your smaller sample to make a conclusion about a larger population. Therefore, the sampling mechanism is very important.

If you want to make a causal statement, such as "if I modify this component, then something else happens," you're going to need experimental data. One type of experimental data is from something like a randomized trial or a randomized study. If you want to make mechanistic types of statements, you need data about all the different components of the system that you're trying to describe.

|  |  |  |
| --- | --- | --- |
| Data Type | Data Source | Outcome/Importance |
| Demographic Data | Basic self-reported demographic characteristics | Characterize the sample |
| Behavioral Assessments | Positive/Negative Affect Schedule (PANAS) Short Stress State Questionnaire (SSSQ) State Trait Anxiety Inventory (STAI) Profile of Mood States (POMS) Dispositional Resilience Scale (DRS) | Measure stress, anxiety, mood, and resilience disposition |
| Executive Function Tests | Iowa Gambling Test; N-Back; Eriksen Flanker | Tests decision making ability, cognitive flexibility, working memory, and response inhibition |
| Performance | Written responses to military test tasks | Adaptability to consistent and inconsistent stimuli in a military context |
| Psychophysiological | Electrocardiography (ECG); Galvanic Skin Response (GSR); Cortisol/??-Amylase | Measure increased and decreased stress, as both an immediate response, and over the study period |
| Self-reports | Written responses after completion of military test tasks | Individual's perceived effectiveness, focus and confidence on the adaptability tasks |

***3. Determining what data you can access***

This is a section on the data we have available to us. Experiment. Explain the participants.

***4. Obtaining the data***

The process for collecting the data in the above table will follow a strict protocol, which is highlighted in this section. The participants in the adaptability study will spend approximately 3 hours in the JHUAPL lab.

* Study participant arrives at JHU/APL testing room.
* The study team will provide an overview of the experiment and study procedures including the psychophysiological measures, sensors and procedures. These instructions will be provided by PowerPoint and study staff will follow an experimenter script across the testing session to assure ensure consistency with each study participant and across data collections.
* The participant will provide written informed consent. Time has been planned to assure that the participant's questions are addressed at this time and throughout the study session.
* A saliva sample will be taken immediately following informed consent and then periodically 7 more times (for a total of 8) throughout the testing.
* A research team member will then instrument the participant with the ECG (2 sensors on belly and collar area) and GSR (two sensors on two fingers of non-dominant hand) sensors. We will use the BioPac sensory system to acquire these data. Baseline measures of the physiological signals will be acquired in order to test for signal integrity and acquire resting baseline values on which data checks may be made.
* The participant will complete initial behavioral assessments including:
  + a demographics questionnaire,
  + the Short Stress State Questionnaire (SSSQ),
  + the State Trait Anxiety Inventory (STAI),
  + the Profile of Mood States (POMS),
  + the Positive and Negative Affect Schedule (PANAS), and
  + the Dispositional Resilience Scale (DRS).
* The participant will complete the neuropsychological tasks:
  + N-back,
  + Ericksen Flanker,
  + IOWA Gambling Task
* The participant will compete the Adaptive Test Tasks.
* When the testing session is completed the ECG and GSR sensors will be removed.

***5. Cleaning the data***

Record the steps of cleaning. Important for reproducibility. What is the code? Determine if the data are good enough. show the str of the final, cleaned data.frame

***6. Exploratory data analysis***

* look at column names
* run head....
* run summary....
* note: need to classify each as NEED TO DEFINE: adaptable and non-adaptable
* initial plots
  + boxplots, transform if needed
  + pairwise comparison to see about correlations
  + cluster analysis, transform if needed....use this to indicate which relationships are interesting

***7. Statistical prediction/modeling***

* should be informed by results of the exploratory analysis
* note: need to tie in the hypothesis stuff on adaptability to something in exploratory data analysis
* must be informed by the question your are interested in solving
* Report measures of uncertainty
* Types:
  + Linear regression to see X/Y correlations
  + tells you the best model to use

***8. Interpretation of results***

* Use appropriate language: describes, correlates with, leads to/causes, predicts, etc.
* give an explaination of what was produced
* interpret coefficients
* Interpret uncertainity

***9. Challenging of results (INTERNAL PROCESS)***

explain why you made decisions on the following: - Question - Data Source - Processing - Analysis - Conclusions.....

***10. Synthesis and write up***

* lead with the question of the study
* TELL A STORY
* summarize the analyses into the story
* don't include every analysis
* order analysis according to the story
* include pretty graphics

***Creating reproducible code***

***Best Practices in Reproducible Research***

1. Don't Do Things By Hand If this chapter could be boiled down to one rule, it might be "Don't do things by hand". What do I mean by that? Here are a few examples that are common, but are bad practice: . Editing spreadsheets of data to "clean it up". Often this is doen to remove outliers, do quality assurance or quality control checks (QA/QC), or validating individual data entries . Editing tables or figures (e.g. rounding, formatting) to make then look better . Downloading data from a web site using a web browser . Moving data around your computer . Splitting or reformatting data files

Often, the motiviation for doing all of the above things is that "We're just going to do this once." The thinking is that if the activity is only going to be done once, it doesn't need to be automated (i.e. programmed into a computer).

But programming a procedure into a computer is not necessarily about automation. It is also about documentation. The problem with things that are done by hand, is that things done by hand need to be precisely documented (this is harder than it sounds). Often, it can very difficult to communicate to someone what was done after the fact. It can be easy to miss a step that "isn't important" when in fact it is.

1. Don't Point And Click Pointing and clicking is obviously related to doing things by hand. Most modern operating systems have a windowing interface that allow you to click on menus that can lead to automated built-in routines. Many data processing and statistical analysis packages have graphical user interfaces (GUIs) that simplify the use of the program, but the actions you take with a GUI can be difficult for others to reproduce because there's not necessarily a log of what was clicked.

Some GUIs for statistical analysis packages produce a log file or script which includes equivalent commands for reproducing the behavior of the GUI, but this is by no means the standard. In general, be careful with data analysis software that is highly interactive. There is often a trade-off between the ease of use of a software package and the tendency to lead to non-reproducible results.

Of course, this doesn't mean that all interactive software. Some software has to be interactive, like text editors or word processors, and that's fine. It's just when the software must be used to conduct data analysis, you must be careful not to be seduced by the ease-of-use of an interactive interface.

1. Teach a Computer The opposite of doing things by hand is teaching a computer to do something. Computers need very precise instructions to accomplish a task so there's no room for ambiguity. This is a Good Thing if your goal is to make your procedures and processes reproducible. If something needs to be done as part of your analysis or investigation, try to teach your computer to do it, even if you only need to do it once. In the end, teaching a computer to do something almost guarantees reproducibilty

***A Philosophy of Data Analysis***

In many ways, according to Peng (2015), the relationship between the Data Analyst and the Principle Investigator is similar to the relationship Filmmakers have with Editors when making a movie. He notes: "In the editing room, the Filmmaker and the Editor can experiment with different versions of different scenes to see which dialogue sounds better, which jokes are funnier,or which scenes are more dramatic. Weak scenes might get dropped, and scenes that are particularly powerful might get extended or re-shot. This "rough cut" of the film is put together quickly so that important decisions can be made about what to pursue further and where to back off. Finer details like color correction or motion graphics might not be implemented at this point. Ultimately, this rough cut will help the director and editor create the "final cut", which is what the audience will ultimately view."

The JHUAPL Data Analyst and JHUAPL Principle Investigator have a similar relationship on the adaptability study. Exploratory data analysis is what occurs in the "editing room" of a research project or any data-based investigation. The JHUAPL team will make a "rough cut" for a data analysis, the purpose of which is very similar to that in the film editing room. The Data Analyst will seek to identify relationships between variables that are particularly interesting or unexpected, checking to see if there is any evidence for or against a stated hypothesis, checking for problems with the collected data. The Data Analyst will always use a subset of data collected by the Principle Investigator, and will also advise the investigator to make critical decisions about what is interesting to follow up on and what probably is not worth pursuing because the data don't provide the evidence.

The JHUAPL Team will employ neuropsychological assessment of executive function to isolate the relationship between specific metrics (inhibition, problem solving and working memory) and adaptability. Participants will also complete a set of computer-based, mission-relevant adaptive test tasks developed to echo elements of the AWG adaptive training program as well as authentic military challenges in order to gauge overall participant adaptable performance. The following table summarizes, for each type of data collected in the Adaptability experiement, the source of the data and why those data are important given the study's purpose.