

Workflow Critique Presentation

PSYC 259 — Principles of Data Science — UC Riverside

Mohammad Dastgheib

February 2022

Today's Presentation

- Part I: Workflow
 - a brief description of my data/project
 - data storage — research personnel — infrastructures
 - data analysis/report packages
- Part II: Critique
 - Efficiency
 - Fidelity
 - Sharing/Reproducibility

a little background — Workflow

Is the Role of Sleep in Memory Consolidation

- For this presentation, I go over the declarative memory data analysis only
- 3 experimental condition (NAP, MED, WAKE) — comparison of memory retention across groups
- Word-pair associate task
 - 60 word-pairs pre-treatment and 60 post-treatment (20 same + 20 novel + 20 recombined)
 - say YES if you've seen the exact pair before

Workflow cont. — other info

- Overall analysis pipeline
 - confusion matrix for their response
 - *Hit, Miss, False Alarm, Correct Rejection*
 - Geometric-mean (G-Mean) as a measure of performance of each participant
 - higher G-Mean \implies superior consolidation
- participants' responses were recorded on paper (series of 0s and 1s)
 - raw values inserted in a Microsoft Excel spreadsheet
 - *Hit, Miss, False Alarm, and Correct Rejection* values calculated by a simple formula
- One graduate student (myself) and two RAs were responsible for collecting data
- Excel sheet imported to the RStudio program for further analyses + write up in RMarkdown

Critique

Efficiency

- data collection improvements
 - employing a simple psychtoolbox code
 - higher accuracy in data collection
 - minimizing error in data entry process
 - facilitation of data transfer to other programs
 - dropping the Microsoft Excel from workflow (?)
- data analysis improvements
 - creating additional cutoff measures to avoid hardcoding


Efficiency cont.

automating the installation of new required packages and preventing reinstalling those already installed.

```
``{r setup, include=FALSE}
#Loading required packages
required_packages <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}

packages <- c("rio","readxl",
"tidyverse","devtools","dplyr","ggplot2","magrittr","Hmisc","psycho","lmerTest","rstanarm","jtools","bayesplot","corrplot")
required_packages(packages)
```


Fidelity

- Sanity check!
 - plotting the distribution of results
- a clear and consistent naming convention
- if required to drop values ± 2 S.D, then clearly annotate
- for the sake of consistency, stick with Tidyverse package
 - admittedly, it may not be 100% possible
- pivoting the datasets
 - at the time, I re-entered the data manually in a Excel spreadsheet and imported in R!
 - 

Sharing/Reproducibility

- my original raw data were posted in OSF and my GitHub repository
- I also wrote the whole dissertation in RMarkdown and uploaded that in my GitHub Repo
 - theoretically, it should be reproducible!
- Something I recently noticed: some packages I used have changed...
 - ideally, using stable packages and avoiding "dev" packages (?)

Thank You

mdast003@ucr.edu