Distributed collaboration for formal modelling

Andy J. Wills

1 Getting started

Getting you set up to start work on distributed collaboration.

1. Install R, RStudio, a PDF viewer, and a spreadsheet application on to your machine.

   (a) R is a free and open-source environment for data analysis. You can download it from `https://cran.r-project.org/index.html`

   (b) RStudio is a front end for R, recommended for most users. You can download it from: `www.rstudio.com/products/rstudio/download/`.

   (c) PDF viewer: Normally already installed. If not, try Evince or Sumatra

   (d) Spreadhseet application: Often already installed. If not, try LibreOffice.

2. Create an RStudio project from the tutorial directory at `www.willslab.org.uk/icps2017`

   (a) Download and unzip *the project directory*

   (b) Open RStudio

   (c) Click on "Project: (None)" (top right)

   (d) Select "new project" and then "existing directory"

   (e) Navigate to the directory you just unzipped and select it

2 Contributing a dataset to a repository

Contributing a dataset requires no programming skills, so it's a good place to start for those new to formal modelling. In these exercises, we go through the process of defining a CIRP, producing a shareable dataset in a standard format, and writing concise documentation for that dataset.

3. Select a CIRP for your research area, and locate a dataset for it

   A CIRP is a Canonical Independently Replicated Phenomenon. It is a specific published experiment that is the best (canonical) representation of a known, independently replicated phenomenon in your field.

   For the purposes of this class, a relatively quick way to find a CIRP is given below. If you need to skip even these steps for reasons of time, the RStudio project you've just created contains some files for a CIRP in category learning (Nosofsky et al., 1994).

   (a) Pick a famous phenomenon in your field,

   (b) Select the most famous (most highly cited?) demonstration of this phenomenon.

   (c) Use a cited reference search (e.g. Web of Science) to confirm it has been independently replicated.

   (d) From this set of replications, pick the one you think is most representative of the phenomenon (and be ready to justify this choice with some brief text later).

4. Create an RData file for your CIRP, using long format.

(a) *Select data from the PDF of the relevant journal article, and copy to the clipboard (use the file* `nosofsky1994.pdf` *provided if you don't have your own, see their Table 1).*

(b) *Paste into your spreadsheet application. (If you can't get copy and paste from a PDF to work on your machine, learn how to do this later, and use the file* `nosof94wide.csv` *provided for now).*

(c) *Use your spreadsheet to re-arrange the data into long format. In long format, each row is one data point, and the columns identify that data point uniquely. For example, the first three data points of the Nosofsky et al. (1994) dataset in long format are:*

type	*block*	*error*
1	*1*	*0.211*
1	*2*	*0.025*
1	*3*	*0.003*

(d) *Save your re-ordered spreadsheet as a CSV file with name* `nosof94long.csv` *into your project directory (use "save as..." and select "CSV" or "Text CSV" from the drop-down list).*

(e) *Use the following command in RStudio to load the CSV file:*
```
nosof94 = read.csv('nosof94long.csv', stringsAsFactors = FALSE)
```

(f) *Now save out in RData format:*
```
save(nosof94, file='nosof94.RData')
```

5. Plan concise documentation for your CIRP.

Work out what you need to say about your CIRP. If you're using Nosofsky et al. (1994), see nosofnotes.txt in the project folder for guidance. CIRP documentation should contain the following information, which you should largely have from earlier in the class.

(a) *Brief (one paragraph) description of the effect, focussing on the IV and on replicated ordinal patterns in the DV*

(b) *Citation of review article that contains full derivation of CIRP. (In the context of this class, we'll assume you haven't yet written such an article, in which case just make up a reference for now).*

(c) *Citations establishing the effect is independently replicated—you should have these from earlier in the class.*

(d) *Reason for selecting this data set as CIRP amongst the replication—again, you should have this earlier in class.*

(e) *Any further brief details you deem relevant about the study.*

6. Create CIRP document in .Rd format

(a) *Open* `blank.Rd` *from your project folder.*

(b) *Fill in the sections as indicated. If a piece of information you produced above does not have a clearly indicated section, put it in the* `details` *section.*

(c) *Save with a filename corresponding to the name of your .RData file (e.g.* `nosof94.Rd` *for* `nosof94.RData`*).*

7. Add your CIRP to a repository

(a) *Locate a dataset repository for your research area that accepts the format you've just created (known as R data package format, see* `http://r-pkgs.had.co.nz/data.html` *). If no such repository is yet available in your research area, talk to me—I can help you start one.*

(b) *For the purposes of the current class, it may be quicker if you use the* `catlearn` *repository, because I administer that one, and can therefore deal with your submissions "live" during class. Don't worry if your CIRP is not from category learning, we can fix that issue after class.*
`http://catlearn.r-forge.r-project.org/`

(c) Read the repository's web pages to discover how to submit data. In the case of `catlearn`*, first-time and occasional contributions are accepted by email—you just email the administrator, attaching the .RData and .Rd files to your email. Regular contributors directly add to the repository from within RStudio.*

8. Celebrate! You just contributed to a distributed collaboration project!

3 Contribute an automated test

For those new to distributed collaboration, writing an automated test is a good "next step" on from submitting a documented dataset. It requires some programming, but generally not as much as implementing a model, or simulating a dataset.

In data repositories that are intended to support distributed collaboration in formal modelling, each CIRP dataset should have an associated automated test function. This is both because such functions provide an unambiguous representation of the reliable findings of the CIRP, and because it allows modellers to effciently test models against a range of phenomena.

9. Write down, in English, what aspects of your dataset a model would have to reproduce to provide an adequate account of its main findings.

   Adequacy can be defined in a number of ways. For the purposes of this class, I suggest you adopt the Ordinal Adequacy Test (OAT) criterion of the `catlearn` *repository; this states that a model simulation is adequate if it reproduces the ordinal pattern of indepenently-replicated significant differences (or well-evidenced differences and equalities, if Bayesian statsitics are available). If you're using the Nosofsky et al. (1994) example, consult the* `nosofnotes.txt` *file for guidance.*

10. Write a computer program that takes a simulated dataset for your CIRP as its input. The program should return "TRUE" when presented with the real dataset, but "FALSE" if one or more of the adequacy conditions you defined above are not met. This can be done in pretty much any computer language but, for the purposes of this class, I suggest you use R.

    (a) Create an empty R function that takes in a dataframe and returns a Boolean; `blankoat.R` *provides an example of how to do this.*

    (b) Fill in that function with a series of commands that convert your English description of adequacy into R. `blankoat.R` *suggests some commands you might find useful. If you need more information on any of these commands, try the help files (e.g. type ?aggregate at the console).*

11. Test your function works properly.

    Test it against both the CIRP data set (it must return "TRUE"), and against various modified datasets that violate your adequacy criteria (it must return "FALSE" in each case).

12. Document your test function in a .Rd file

    This is a similar task to documenting your CIRP in a .Rd file, see `blankoat.Rd` *for a template.*

13. Add your test function to the same repository as your CIRP.

    See guidance for adding your CIRP to a repository.

14. Celebrate again! You've completed an intermediate-level contribution to distributed collaboration.