R Session 7: Reproducible Analysis

Steps toward enabling duplication of results by others

William A. Cooper

Research Aviation Facility, Earth Observing Laboratory
National Center for Atmospheric Research

GOALS

Precept: The analysis can be duplicated by someone else using provided information.

Suggested information to provide:

1. The text, with documentation of all the analysis steps, results, and interpretation. (The interpretation, of course, may be different when evaluated by someone else.)
2. The code used. This may be in terms of a program for a specific language.
3. Enough information on the underlying language (version, operating system, etc) that someone else can use the same code interpreter if necessary.
4. Locations of data files used, if in sustained archives, or copies of the data sufficient to reproduce the results.

THE TRADITIONAL WAY

A possible approach

- Write and post or publish the text document.
- In that document, reference the program and data set used.
- Make the program available in some repository.
- Ensure the data set is archived where it is accessible.

Some dangers:

- Often, program steps are scattered and hard to assemble, with different steps used to generate plots, manipulate data, perform fits, construct derived data, use multiple and supplementary data sets, etc.
- Repositories may have short lifetimes, esp. as used for programs.
- Reproducibility often means following a tortuous path.
- Data sets often are revised, or archives disappear, and program steps may be revised without documentation.

A MORE STRUCTURED MODEL:

Using R with knitr (esp. with RStudio):

1. Embed the text and code in the same document. Use 'knitr' functions to reference results from the code in the text to ensure consistency.
2. Package the result with these components:
   (a) The "Rnw" or "R-html" or "Markdown" file that generates the text and runs the code. (This is possible using some other programs other than R also, and can work with LaTeX or html code or some other formats also.)
   (b) The "publication" – usually PDF or html.
   (c) Any data that should be included in the package. Use DOIs where possible.
   (d) A file documenting the version of various programs or computer systems used.
3. Archive this package in a repository with long expected lifetime and appropriate accessibility.

A SIMPLE EXAMPLE:

Attack calibration based on a speed run

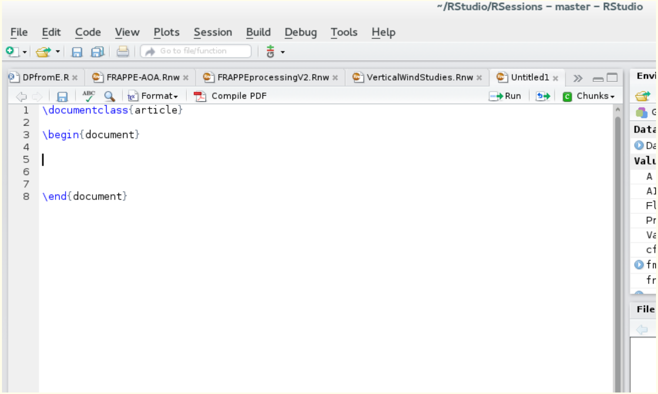
1. Describe the algorithm
2. Get the data needed
3. Do the fit
4. Construct a plot
5. Discuss the result
6. Embed reproducibility information directly in the file.

THE "Rnw" file: (LaTeX approach)

1. In RStudio, select 'File->New File->R Sweave
2. Embed documentation, either directly or generated by Word or Libreoffice or lyx, etc.
3. Embed the code to do the calculation.
4. Reference the results of the calculation in the text.
5. Add an appendix with reproducibility information.

FRAMEWORK step 1:

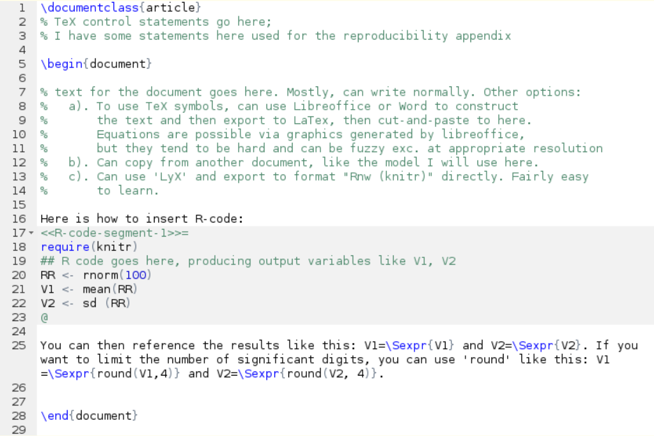
File -> New -> Sweave



```
\documentclass{article}

\begin{document}



\end{document}
```

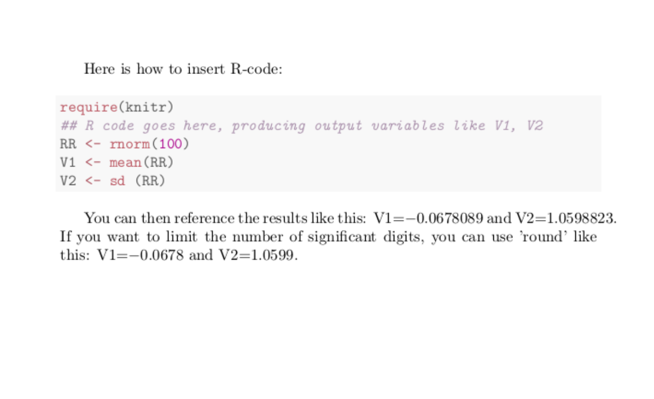
FRAMEWORK step 2:

Fill in the sections as follows:



```
\documentclass{article}
% TeX control statements go here;
% I have some statements here used for the reproducibility appendix

\begin{document}

% text for the document goes here. Mostly, can write normally. Other options:
%   a). To use TeX symbols, can use Libreoffice or Word to construct
%       the text and then export to LaTex, then cut-and-paste to here.
%       Equations are possible via graphics generated by libreoffice,
%       but they tend to be hard and can be fuzzy exc. at appropriate resolution
%   b). Can copy from another document, like the model I will use here.
%   c). Can use 'LyX' and export to format "Rnw (knitr)" directly. Fairly easy
%       to learn.

Here is how to insert R-code:
<<R-code-segment-1>>=
require(knitr)
## R code goes here, producing output variables like V1, V2
RR <- rnorm(100)
V1 <- mean(RR)
V2 <- sd (RR)
@

You can then reference the results like this: V1=\Sexpr{V1} and V2=\Sexpr{V2}. If you want to limit the number of significant digits, you can use 'round' like this: V1=\Sexpr{round(V1,4)} and V2=\Sexpr{round(V2, 4)}.


\end{document}

```

OUTPUT SO FAR:



Here is how to insert R-code:

```
require(knitr)
## R code goes here, producing output variables like V1, V2
RR <- rnorm(100)
V1 <- mean(RR)
V2 <- sd (RR)
```

You can then reference the results like this: V1=−0.0678089 and V2=1.0598823. If you want to limit the number of significant digits, you can use ’round’ like this: V1=−0.0678 and V2=1.0599.

REALISTIC EXAMPLE: AOA CAL FOR FRAPPE

See program FRAPPE-AOA.Rnw, noting:

1. top entries – just for Reproducibility section, can just copy
2. Insertion of title. author, date
3. Text, including equations and symbols
4. inclusion of R code. Can insert via "Code->Insert Chunk"
   (a) "echo" shows code in document
   (b) "include" shows any printed results or plots in the document
   (c) These can be placed so as to get output where desired in the document
5. Reference R results via the \Sexpr expression in the text.
6. See the PDF version, generated by "Compile PDF" in RStudio
7. Note especially the appended "Reproducibility" section.

MORE ABOUT THE REPRODUCIBILITY SECTION

Structured to remind of needed information

1. Lists what the project is called. (Not the same as an RStudio project, but it could be).
2. Defines the package to be archived and the name of the program
3. Also identifies the original data and, in cases like this where the data file is small (here, 12K), includes the data file in the zip file for archiving.
4. Provides the name of the archive where this zip file can be accessed.
5. Lists, via embedded code segments that look up this info from the netCDF file, some of the relevant calibrations used for processing.

OTHER WAYS (ESP. AVOIDING LATEX):

Still using knitr: Rmd document using Markdown

- Similar in concept, but code-segment identifiers are different and the output is normally an HTML document but can be PDF or Word/Libreoffice.
- Many Markdown items like \alpha are LaTeX-like but easier to learn.
- FRAPPE-AOA.Rmd shows the same document as before but in Markdown.
- Accepts HTML code as well as Markdown.
- Good alternative to using LaTeX
- You can also generate presentations in Markdown.

FURTHER INFORMATION RE KNITR

Lots of capability; commensurate learning curve to use fully

- "Dynamic Documents with R and knitr" by Yihui Xie: The complete reference by the creator of knitr.
- "Reproducible Research with R and RStudio" by Christopher Gandrud, with lots of advice on the *process* of conducting reproducible research
- Incorporating tables requires some additional work with LaTeX format
- For serious users, LyX is a great tool to use for generating Rnw files. It is adapted to work with knitr and R, and it has good presentation capabilities. All these RSession presentations were generated with LyX using the 'beamer' package, and you can find the lyx files on the archive locations including on tikal.