Sweave as a tool for reproducible quantitative analysis reporting

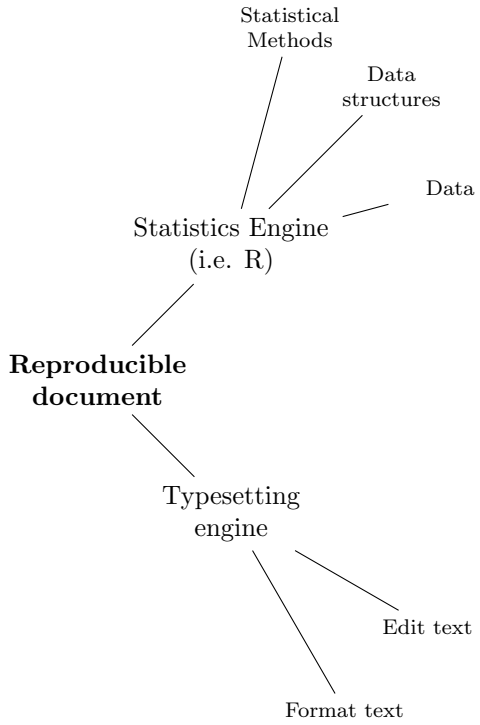
Kris Wright

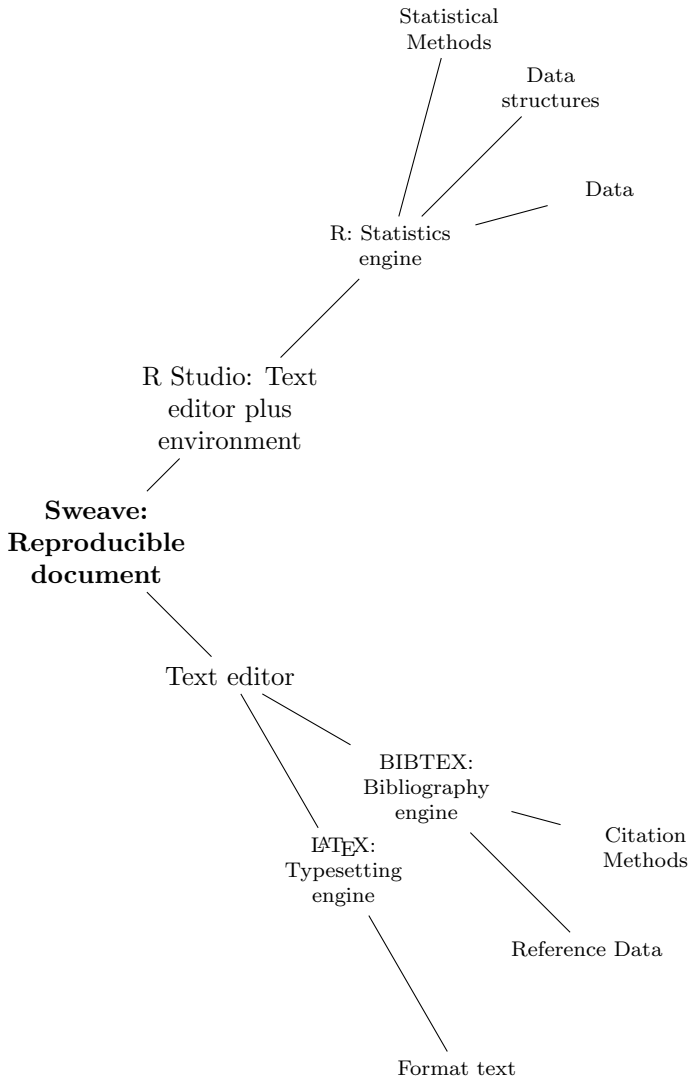
April 12, 2013

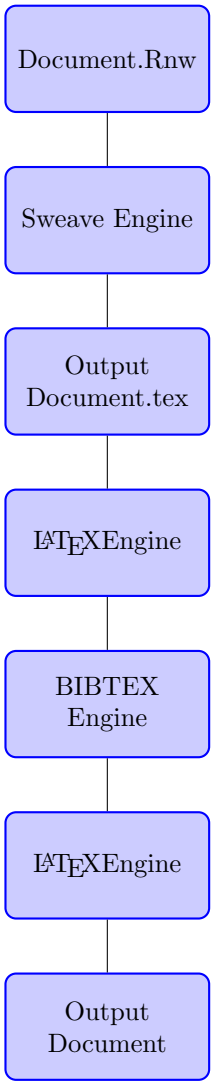
1 The problems

- A journal or another stakeholder wants high visibility into the quantitative methods and analysis of a paper
- A researcher wants to reproduce results 10 years later
- A researcher would rather invest additional time up front to avoid editing errors and time later
- A student finds out the day before a highly quantitative paper is due that some numbers were mistyped in the data file

2 Anatomy of reproducible quantitative analysis reporting




Statistical Methods
Data structures
Data
R: Statistics engine
R Studio: Text editor plus environment
Sweave: Reproducible document
Text editor
BIBTEX: Bibliography engine
Citation Methods
LaTeX: Typesetting engine
Reference Data
Format text


Document.Rnw
Sweave Engine
Output
Document.tex
LaTeXEngine
BIBTEX
Engine
LaTeXEngine
Output
Document

3 Options for reproducible quantitative analysis reporting

- MS Word
- Open Office
- LaTeX

3.0.1 Word advantages / disadvantages

- Can't use Sweave with MS Word
- There is code for calling R functions in excel. Therefore, it is most likely possible to call R using excel objects in MS Word
- Word is known as a WYSIWYG?
- Most likely already "know" MS Word?

3.0.2 Open Office advantages / disadvantages

- Can use Sweave with Open Office! (as well as HTML)
- Open Office has some similarities to MS Word.
- Freely Available!
- Is not LaTeX

3.0.3 LaTeXadvantages

- Easier!
- Active Community!
- Freely Available!
- Is ordinary text (no need to worry about file formats changing years later!)
- Faster! (no GUI to slow down!)
- Disadvantages? (maybe installing it and initial learning curves...)

4 LaTeX

4.1 Basic Structure

4.1.1 Code

```
\documentclass{article}

\begin{document}

Hi!  I am a simple \LaTeX document.

\end{document}
```

4.1.2 Output

Hi! I am a simple LaTeXdocument.

4.2 Extend functionality with packages

4.2.1 Code

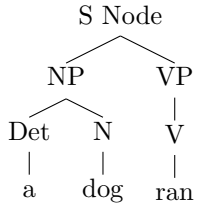
```
\documentclass{article}

\usepackage{qtree}
\begin{document}

\Tree[.{S Node} [.NP [.Det a ] [.N dog ] ] [.VP [.V ran ] ] ]

\end{document}
```

4.2.2 Output



4.3 Add bibliographic references

4.3.1 Code

```
\documentclass{article}
\usepackage{apacite}

\begin{document}
```

```
According to \citeA{sweavehome}, there is a lot of significant functionality.

\bibliographystyle{apacite}
\bibliography{sweavepresentation_Friday}{}
\end{document}
```

4.3.2 Output

According to Leisch (2013), there is a lot of significant functionality.

References

Leisch, F. (2013, 04). *Lmu - prof. friedrich leisch.* Retrieved 04/11/2013, from http://www.statistik.lmu.de/ leisch/Sweave/

4.4 LaTeXpart of Sweave

4.4.1 Code

```
\documentclass{article}

\usepackage{Sweave}

\begin{document}

Next is the result of an actual R calculation:

<<echo=FALSE, results=tex >>=

a_variable = 1 + 1
a_variable

@

\par
\noindent Cool!

\end{document}
```

4.4.2 Output

Next is the result of an actual R calculation: [1] 2
Cool!

5 Sweave Chunks

Sweave code is intermixed with chunks that look like the following:

$<< echo = FALSE, results = hide >>=$
Sweave Code goes here
the # sign is a comment in R, and thus Sweave
@

Figure 1: Basic sweave chunk

R code chunks can be reused if given a label. This is especially nice when making graphs.

$<< label = alabel, echo = FALSE, results = hide >>=$
Sweave Code goes here
@
$<< echo = FALSE, results = hide >>=$
<<alabel>>
@

Figure 2: Sweave chunk reuse

Other options for echo is TRUE and FALSE results is verbatim, tex and hide

5.1 Add some numbers together and print

5.1.1 Display all : $echo = TRUE, results = verbatim$

```
> a_variable = 1 + 1
> a_variable
```

```
[1] 2
```

5.1.2 Display results only: $echo = FALSE, results = verbatim$

```
[1] 2
```

5.1.3 Display results only: $echo = FALSE, results = tex$

[1] 2

5.1.4 No display: $echo = FALSE, results = hide$

The variable from the Sweave code can be assessed using the
Sexpr command to give it's value = 2.

5.2 Example

5.2.1 dataframe

```
> testframe <- data.frame(speakertype="speaker", subj=1, itemN=1, ptype="C", plurality="P",
> testframe <- testframe[testframe$subj == 500,];
> for (i in 1:2) {
+ testframe <- rbind(testframe, data.frame(speakertype="speaker", subj=i, itemN=1:3,ptype="C
+ }
> for (i in 3:4) {
+ testframe <- rbind(testframe, data.frame(speakertype="speaker2", subj=i, itemN=1:3,ptype='
+ }
```

5.2.2 xtable with a dataframe

```
> # Will use the xtable library for this paper.  Though I need to do some
> #  Formatting on it next time.
> library(xtable);

>  xtable(testframe)
```

	speakertype	subj	itemN	ptype	plurality	rt
1	speaker	1	1	C	P	75.03
2	speaker	1	2	C	P	82.07
3	speaker	1	3	C	P	86.15
4	speaker	2	1	C	P	86.32
5	speaker	2	2	C	P	90.12
6	speaker	2	3	C	P	85.14
7	speaker2	3	1	I	S	81.41
8	speaker2	3	2	I	S	73.92
9	speaker2	3	3	I	S	91.19
10	speaker2	4	1	I	S	101.05
11	speaker2	4	2	I	S	92.47
12	speaker2	4	3	I	S	70.88

5.2.3 results=verbatim vs results=tex

```
> data(tli)
> tli_aov <- aov(tlimth ~ sex * ethnicty + Error(grade/(sex+ethnicty)), data=tli)
> sum_tli <- summary(tli_aov)
```

results=verbatim

```
> sum_tli
```

```
Error: grade
    Df Sum Sq Mean Sq
sex  1  362.2   362.2

Error: grade:sex
    Df Sum Sq Mean Sq
sex  1  39.24   39.24

Error: grade:ethnicty
         Df Sum Sq Mean Sq
sex       1  540.2   540.2
ethnicty  2 1294.0   647.0

Error: Within
             Df Sum Sq Mean Sq F value Pr(>F)
sex           1     91   91.31   0.438  0.510
ethnicty      3    513  170.92   0.820  0.486
sex:ethnicty  2    246  123.22   0.591  0.556
Residuals    88  18340  208.41
```

results=tex

```
> sum_tli
```

Error: grade Df Sum Sq Mean Sq sex 1 362.2 362.2
Error: grade:sex Df Sum Sq Mean Sq sex 1 39.24 39.24
Error: grade:ethnicty Df Sum Sq Mean Sq sex 1 540.2 540.2 ethnicty 2 1294.0 647.0
Error: Within Df Sum Sq Mean Sq F value Pr(>F) sex 1 91 91.31 0.438 0.510 ethnicty 3 513 170.92 0.820 0.486 sex:ethnicty 2 246 123.22 0.591 0.556 Residuals 88 18340 208.41

5.2.4 Accessing parts of a dataframe in R

```
> sum_tli[[length(sum_tli)]][[1]]

             Df  Sum Sq Mean Sq F value Pr(>F)
sex           1    91.3  91.309  0.4381 0.5098
ethnicty      3   512.8 170.922  0.8201 0.4862
sex:ethnicty  2   246.4 123.224  0.5913 0.5558
Residuals    88 18339.9 208.408

> sum_tli[[length(sum_tli)]][[1]][[3]]

[1]  91.30864 170.92184 123.22429 208.40781

> sum_tli[[length(sum_tli)]][[1]][[3]]
```

```
[1]  91.30864 170.92184 123.22429 208.40781
```

```
> sum_tli[[length(sum_tli)]][[1]]
```

```
             Df  Sum Sq Mean Sq F value Pr(>F)
sex           1    91.3  91.309  0.4381 0.5098
ethnicty      3   512.8 170.922  0.8201 0.4862
sex:ethnicty  2   246.4 123.224  0.5913 0.5558
Residuals    88 18339.9 208.408
```

```
> sum_tli[[length(sum_tli)]][[1]]
```

```
             Df  Sum Sq Mean Sq F value Pr(>F)
sex           1    91.3  91.309  0.4381 0.5098
ethnicty      3   512.8 170.922  0.8201 0.4862
sex:ethnicty  2   246.4 123.224  0.5913 0.5558
Residuals    88 18339.9 208.408
```

```
> sum_tli_2 <- summary(aov(tlimth ~ sex * ethnicty, data=tli))
> sum_tli[[length(sum_tli)]][[1]][[3]]
```

```
[1]  91.30864 170.92184 123.22429 208.40781
```

```
> sum_tli[[length(sum_tli)]][[1]][[5]]
```

```
[1] 0.5097585 0.4861734 0.5558105        NA
```

```
> sum_tli[[length(sum_tli)]][[1]][[1]][[1]]
```

```
[1] 1
```

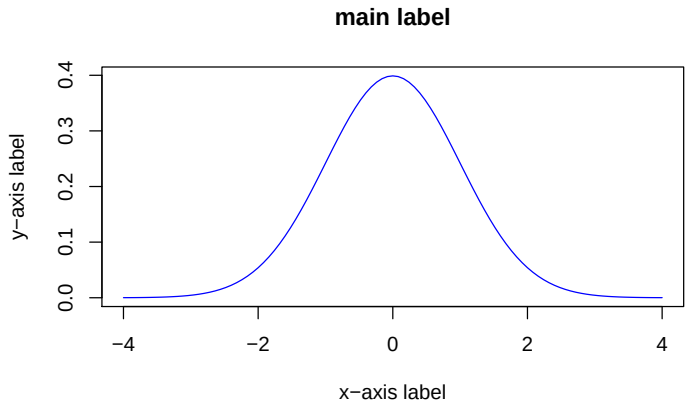
5.3 xtable with a matrix

```
> mat <- round(matrix(c(0.9, 0.89, 200, 0.045, 2.0), c(1, 5)), 4);
> print(xtable(mat))
```

	1	2	3	4	5
1	0.90	0.89	200.00	0.04	2.00

5.4 Include a graph

$<< fig = true, echo = FALSE, results = hide, width = 6, height = 3.8 >>=$
plot(function(x)dnorm(x), -4, 4, col=c("blue"), xlab = "x-axis label", ylab="y-axis label", main="main label")
@



6 Converting to other datatypes in R

results=verbatim

to a list:

```
> datfram2 <- data.frame(sum_tli[[length(sum_tli)]][[1]])
> datfram2

             Df      Sum.Sq   Mean.Sq   F.value    Pr..F.
sex           1    91.30864  91.30864 0.4381248 0.5097585
ethnicty      3   512.76552 170.92184 0.8201316 0.4861734
sex:ethnicty  2   246.44858 123.22429 0.5912652 0.5558105
Residuals    88 18339.88734 208.40781        NA        NA

> mylist <- as.list(datfram2[5])
> mylist

$Pr..F.
[1] 0.5097585 0.4861734 0.5558105        NA
```

to a matrix:

```
> mat2 <- as.matrix(sum_tli[[length(sum_tli)]][[1]])
> mat2

             Df      Sum Sq   Mean Sq   F value    Pr(>F)
sex           1    91.30864  91.30864 0.4381248 0.5097585
ethnicty      3   512.76552 170.92184 0.8201316 0.4861734
sex:ethnicty  2   246.44858 123.22429 0.5912652 0.5558105
Residuals    88 18339.88734 208.40781        NA        NA
```