Writing reproducible reports

knitr with R Markdown

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

Statisticians write a lot of reports, describing the results of data analyses. It's best if such reports are fully reproducible: that the data and code are available, and that there's a clear and automatic path from data and code to the final report.

knitr is ideal for this effort. It's a system for combining code and text into a single document. Process the document, and the code is replaced with the results and figures that it generates.

I've found it most efficient to produce informal analysis reports as web pages. Markdown is a system for writing simple, readable text, with the sort of marks that you might use in an email message, that gets converted to nicely formatted html-based web pages.

My goal in this lecture is to show you how to use knitr with R Markdown (a variant of Markdown) to make such reproducible reports, and to convince you that this is the way that you should be constructing such analysis reports.

I'd originally planned to also cover knitr with AsciiDoc, but I decided to drop it; it's best to focus on Markdown.

How many simulation replicates?

- To estimate power?
- To estimate a p-value?
- To estimate some other quantity?

First, a quick note from last time: in computer simulations to estimate power, or to estimate a p-value or some other quantity, how many simulation replicates should you use?

For power, consider that the estimated count is binomial(n, p) where n is the number of replicates and p is the true power. The estimated power will have standard error $\sqrt{p(1-p)/n}$. You maybe want to choose n so that this is like 1%.

For estimating a p-value, you're in about the same situation. Your estimated p-value has SE $\sqrt{p(1-p)/n}$ so choose n to make this as small as you'd like for the sort of p-values you're expecting.

For other parameters, it depends on how variable they are. You can always start with 100 and use those results to get an estimate of how many more you might need.

Data analysis reports

- Figures/tables + email
- Static Word document
- LaTeX + R $\rightarrow$ PDF
- R Markdown = knitr + Markdown $\rightarrow$ Web page

Statisticians write a lot of reports. You do a bunch of analyses, create a bunch of figures and tables, and you want to describe what you've done to a collaborator.

When I was first starting out, I'd create a bunch of figures and tables and email them to my collaborator with a description of the findings in the body of the email. That was cumbersome for me and for the collaborator. ("Which figure are we talking about, again?")

I moved towards writing formal reports in LaTeX and sending my collaborator a PDF. But that was a lot of work, and if I later wanted to re-run things (e.g., if additional data were added), it was a real hassle.

It's also a pain to deal with page breaks in PDF documents.

Web pages, produced with knitr and Markdown, are ideal. You can make super-tall multi-panel figures that show the full details, without worrying page breaks. And hyperlinks are more convenient, too.

What if the data change?

What if you used the wrong version of the data?

If data are added, will it be easy to go back and re-do your analyses, or is there a lot of copying-and-pasting and editing to be done?

I usually start an analysis report with a summary of the experiment, scientific questions, and the data. Recently, a collaborator noticed that I'd used an old version of the data. (I'd cited sample sizes, and so he could see that I didn't have the full set.)

He said, "I'm really sorry you did all that work on the incomplete dataset."

But actually, it didn't take long to find the right file, and the revised analysis was derived instantaneously, as I'd used knitr.

knitr in a knutshell

kbroman.org/knitr_knutshell

rmarkdown.rstudio.com

I wrote a short tutorial on knitr, covering a bit more than I'll cover in this lecture.

I'd be glad for suggestions, corrections, or questions.

knitr code chunks

Input to knitr:

```
We see that this is an intercross with `r nind(sug)`
individuals. There are `r nphe(sug)` phenotypes, and genotype
data at `r totmar(sug)` markers across the `r nchr(sug)`
autosomes.  The genotype data is quite complete.

```{r summary_plot, fig.height=8}
plot(sug)
```
```

Output from knitr:

```
We see that this is an intercross with 163
individuals. There are 6 phenotypes, and genotype
data at 93 markers across the 19
autosomes.  The genotype data is quite complete.

```r
plot(sug)
```

![plot of chunk summary_plot](RmdFigs/summary_plot.png)
```

The basic idea in knitr is that your regular text document will be interrupted by chunks of code delimited in a special way.

This example is with R Markdown.

There are in-line bits of code indicated with backticks. When the document is processed by knitr, they'll be evaluated and replaced by the result.

Larger code chunks with three backticks. This one will produce a plot. When processed by knitr, an image file will be created and a link to the image will be inserted at that location.

In knitr, different types of text have different ways of delimiting code chunks, because it's basically going to do a search-and-replace and depending on the form of text, different patterns will be easier to find.

html

```
<!DOCTYPE html>
<html>
<head>
  <meta charset=utf-8"/>
  <title>Example html file</title>
</head>

<body>
<h1>Markdown example</h1>

<p>Use a bit of <strong>bold</strong> or <em>italics</em>. Use
backticks to indicate <code>code</code> that will be rendered
in monospace.</p>

<ul>
<li>This is part of a list</li>
<li>another item</li>
</ul>

</body>
</html>
```


It's helpful to know a bit of html, which is the markup language that web pages are written in. html really isn't that hard; it's just cumbersome.

An html document contains pairs of tags to indicate content, like `<h1>` and `</h1>` to indicate that the enclosed text is a "level one header", or `<em>` and `</em>` to indicate emphasis (generally italics). A web browser will parse the html tags and render the web page, often using a cascading style sheet (CSS) to define the precise style of the different elements.

Note that there are six levels of headers, with tags `<h1>`, `<h2>`, `<h3>`, …, `<h6>`. Think of these as the title, section, subsection, sub-subsection, …

CSS

```
ul,ol {
  margin: 0 0 0 35px;
}

a {
  color: purple;
  text-decoration: none;
  background-color: transparent;
}

a:hover
{
  color: purple;
  background: #CAFFFF;
}
```

[Example]

I don't really want to talk about CSS, but I thought I should at least acknowledge its existence.

CSS is really important for defining how your document will appear. Much of the time, you just want to find someone else's CSS document that is satisfactory to you.

Markdown

````
# Markdown example

Use a bit of **bold** or _italics_. Use backticks to indicate
`code` that will be rendered in monospace.

- This is part of a list
- another item

Include blocks of code using three backticks:

```
x <- rnorm(100)
```

Or indent four spaces:

    mean(x)
    sd(x)

And it's easy to create links, like to
[Markdown](http://daringfireball.net/projects/markdown/).
````


Markdown is a system for writing simple, readable text that is easily converted into html. The reason it's useful to know a bit of html is that then you have a better idea how the final product will look. (Plus, if you want to get fancy, you can just insert a bit of html within the Markdown document.)

Markdown is just a system of marks that will get searched-and- replaced to create an html document. A big advantage of the Markdown marks is that the source document is much like what you might write in an email, and so it's much more human-readable.

Github (which we'll talk about next week) automatically renders Markdown files as html, and you can use Markdown for ReadMe files. And the website for this course is mostly in Markdown.

R Markdown

- R Markdown is a variant of Markdown, developed at RStudio.com
- Markdown + knitr + extras
- A few extra marks
- LaTeX equations
- Bundle images into the final html file

R Markdown is a variant of Markdown developed by the folks at RStudio.

It's Markdown with knitr code chunks, but there are a number of added features, most importantly the ability to use LaTeX equations.

YAML header

```
---
title: "knitr/R Markdown example"
author: "Karl Broman"
date: "28 January 2015"
output: html_document
---
```

```
---
title: "Another knitr/R Markdown example"
author: "[Karl Broman](https://kbroman.org)"
date: "`r Sys.Date()`"
output: word_document
---
```


At the top of your Rmd file, it's best to include a header like the above examples. (YAML is a simple text-based format for specifying data, sort of like JSON but more human-readable.)

You don't have to include any of these things, but it's good to at least specify `output:` (which can be also be `pdf_document`). There are a lot more options; see `rmarkdown.rstudio.com`.

Note my use of a hyperlink and some R code in the second example. These will carry over to the final document.

`?rmarkdown::html_document`

- `toc_float`
- `toc_depth`
- `code_folding`
- `theme`
- `df_print`

There are a lot of great options that you can shove into that yaml header. Here are some of my favorite ones. You can learn about others by looking at the help file for `html_document`.

Code chunks, again

````
```{r knitr_options, include=FALSE}
knitr::opts_chunk$set(fig.width=12, fig.height=4,
                      fig.path='Figs/', warning=FALSE,
                      message=FALSE)
set.seed(53079239)
```

### Preliminaries

Load the R/qtl package using the `library` function:

```{r load_qtl}
library(qtl)
```

To get help on the read.cross function in R, type the
following:

```{r help, eval=FALSE}
?read.cross
```
````


A couple of additional points about code chunks.

You can (and should) assign names to the code chunks. It will make it easier to fix errors, and figure files will be named based on the name of the chunk that produces them.

Code chunks can also have options, like `include=FALSE` and `eval=FALSE`. And you can define global options, which will apply to all subsequent chunks.

Chunk options

`echo=FALSE`	Don't include the code
`results="hide"`	Don't include the output
`include=FALSE`	Don't show code or output
`eval=FALSE`	Don't evaluate the code at all
`warning=FALSE`	Don't show R warnings
`message=FALSE`	Don't show R messages
`fig.width=#`	Width of figure
`fig.height=#`	Height of figure
`fig.path="Figs/"`	Path for figure files

There are lots of chunk options.

These are the chunk options that I use most, but there are lots more. Each should be valid R code, and can be basically any valid R code, so you can get pretty fancy.

The ending slash in `fig.path` is important, as this is just pasted to the front of the figure file names. If not included, the figures would be in the main directory but with names starting with "`Figs`".

Global chunk options

````
```{r knitr_options, include=FALSE}
knitr::opts_chunk$set(fig.width=12, fig.height=4,
                      fig.path='Figs/', warning=FALSE,
                      message=FALSE, include=FALSE,
                      echo=FALSE)
set.seed(53079239)
```

```{r make_plot, fig.width=8, include=TRUE}
x <- rnorm(100)
y <- 2*x + rnorm(100)
plot(x, y)
```
````

- Use global chunk options rather than repeat the same options over and over.
- You can override the global values in specific chunks.

I'll often use `include=FALSE` and `echo=FALSE` in a report to a collaborator, as they won't want to see the code and raw results. I'll then use `include=TRUE` for the figure chunks.

And I'll set some default choice for figure heights and widths but then adjust them a bit in particular figures.

You may need to include `\library(knitr)` before the `opts_chunk$set()` (for example, within RStudio).

Package options

```
```{r package_options, include=FALSE}
knitr::opts_knit$set(progress = TRUE, verbose = TRUE)
```
```

- It's easy to confuse global chunk options with package options.
- I've not used package options.
- So focus on `opts_chunk$set()` not `opts_knit$set()`.

If you are doing something fancy, you may need knitr package options, but I've not used them.

I've gotten confused about them, though: `opts_chunk$set` vs. `opts_knit$set`.

In-line code

```
We see that this is an intercross with `r nind(sug)`
individuals. There are `r nphe(sug)` phenotypes, and genotype
data at `r totmar(sug)` markers across the `r nchr(sug)`
autosomes.  The genotype data is quite complete.
```

- Each bit of in-line code needs to be within one line; they can't span across lines.
- I'll often precede a paragraph with a code chunk with `include=FALSE`, defining various variables, to simplify the in-line code.
- Never hard-code a result or summary statistic again!

In-line code to insert summary statistics and such is a key feature of knitr.

Even if you wanted the code for your figures or data analysis to be separate, you'd still want to make use of this feature.

Remember my anecdote earlier in this lecture: if I hadn't mentioned sample sizes, my collaborator wouldn't have noticed that I was using an old version of the data.

Python in R Markdown

You can have python code chunks in R Markdown. And information is remembered between chunks.

````
```{python define_something}
x = [2, 3, 5, 7, 9, 11, 13, 17]
```

```{python list_comprehension}
y = [v*2 for v in x]
```
````

It seems like you can't use python in-line. But if load the package 'reticulate', you can get access to python objects with R code.

```
  The first value in `x` is `r py$x[1]`, while the first value in `y` is
  `r py$y[1]`.
```

More at rstudio.github.io/reticulate/

You can have python code chunks in an R Markdown document, and the objects you create have permanence between chunks: there's a python environment running continuously the way R does.

And with the reticulate package, you can get access to R objects from python and vice versa. But it can be a little ugly.

One limitation is that you can't seem to use python in-line the way you can with R. All I can see to do, if you want a bit of in-line code to show some statistics, is to use R in-line and refer to the python objects through this rather ugly `py` object which contains all of the python objects as a big named list.

Rounding

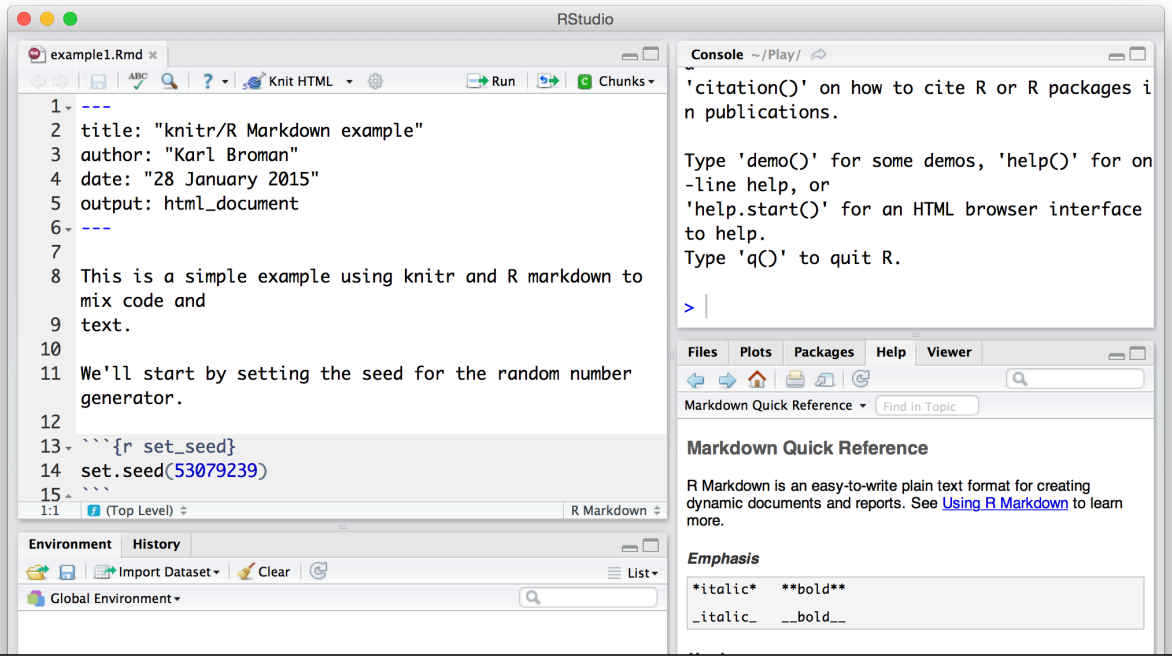
- `cor(x,y)` might produce `0.8992877`, but I want `0.90`.
- `round(cor(x,y), 2)`, would give `0.9`, but I want `0.90`.
- You could use `sprintf("%.2f", cor(x,y))`, but `sprintf("%.2f", -0.001)` gives `-0.00`.
- Use the `myround` function in my R/broman package.
- `myround(cor(x,y), 2)` solves both issues.

I'm very particular about rounding. You should be too.

If you're a C programmer, sprintf seems natural. No one else agrees.

The R/broman package is on both github and CRAN.

R Markdown → html, in RStudio



The easiest way to convert an R Markdown file to html is with RStudio.

Open the R Markdown file in R Studio and click the "Knit HTML" button (with the ball of yarn and knitting needle).

Note the little button with a question mark. Click that, and you'll get the "Markdown Quick Reference."

What actually happens: The `knit` function in the knitr package processes all of the code chunks and in-line code and creates a Markdown file and possibly a bunch of figure files. The Markdown file (and any figure files) are sent to Pandoc, which converts them to an HTML file, with embedded figures.

RStudio is especially useful when you're first learning R Markdown and knitr, as it's easy to create and view the corresponding html file, and you have access to that Markdown Quick Reference.

R Markdown → html, in R

```
> library(rmarkdown)
> render("knitr_example.Rmd")
```

```
> rmarkdown::render("knitr_example.Rmd")
```


When you click the "Knit HTML" button in RStudio, what it actually does is run `rmarkdown::render()`, which in turn calls `knitr::knit()` and then runs pandoc.

You can do the same thing directly, in R. You do miss out on the immediate preview of the result.

R Markdown → html, GNU make

```
knitr_example.html: knitr_example.Rmd
    R -e "rmarkdown::render('knitr_example.Rmd')"
```


I prefer to do this from the command-line, using a Makefile. Then it's more obvious what's happening.

In Windows, it's important that the double-quotes are on the outside and the single-quotes are on the inside.

Need pandoc in your `PATH`

RStudio includes pandoc; you just need to add the relevant directory to your `PATH`.

Mac:

```
/Applications/RStudio.app/Contents/MacOS/pandoc
```

Windows:

```
"c:\Program Files\RStudio\bin\pandoc"
```


To use the rmarkdown package from the command line, you need access to pandoc. But if you've installed RStudio (and I highly recommend that you do), you don't need to do a separate install, as pandoc is included with RStudio.

You just need to add the relevant directory (listed above) to your `PATH`, for example in your `~/.bash_profile` file.

At the command line, type `type pandoc` or `pandoc --version` to check that it's available.

Reproducible knitr documents

- Don't use absolute paths like `~/Data/blah.csv`
- Keep all of the code and data in one directory (and its subdirectories)
- If you must use absolute paths, define the various directories with variables at the top of your document.
- Use `R --vanilla` or perhaps

```
R --no-save --no-restore --no-init-file --no-site-file
```

- Use GNU make to document the construction of the final product (tell future users what to do)
- Include a final chunk with `getwd()` and `devtools::session_info()`.
- For simulations, use `set.seed` in your first chunk.

That you've used knitr doesn't mean the work is really reproducible. The source and data need to be available to others, they need to know what packages were used and how to compile it, and then they need to be able to compile it on their system.

The complicated alternative to `R --vanilla` is if you want to still load `~/.Renviron`, for example, to define `R_LIBS`.

If you use `set.seed` at the top of the document, it should be that the random aspects will give exactly the same results. I'll use
`runif(1, 0, 10^8)` and then paste that big number within `set.seed()`.

Two anecdotes: The github repository for the Reproducible Research with R and R Studio book uses some absolute paths that basically make it not reproducible.

Earn et al. (2014) Proc Roy Soc B 281(1778):20132570 has a really nice supplement, written with knitr. But it says, "The source code is available upon request." It's not really reproducible, then.

Controlling figures

```
```{r test_figure, dev.args=list(pointsize=18)}
x <- rnorm(100)
y <- 2*x + rnorm(100)
plot(x,y)
```
```

- The default is for knitr/R Markdown is to use the `png()` graphics device.
- Use another graphics device with the chunk option `dev`.
- Pass arguments to the graphics device via the chunk option `dev.args`.

Graphics in knitr are super easy. For the most part, you don't have to do anything! If a code chunk produces a figure, it will be inserted.

But depending on the type of figure, you might want to try different graphics devices. And sometimes you want to pass arguments to the graphics device.

Yesterday (6 Feb 2014), to change the size of axis labels, you couldn't just use the pointsize device argument; you'd also need to use something like `par(cex.lab=1.5)`. But I posted a question about it on StackOverflow, and Yihui Xie responded and then immediately fixed the problem. I used a bit of twitter in there too, to get his attention.

To download and install the development version of knitr, you can use the `install_github` function in Hadley Wickham's devtools package. Use `install.packages("devtools")` if you don't already have it installed. Then `library(devtools)` and `install_github("yihui/knitr")`.

Tables

````
```{r kable}
x <- rnorm(100)
y <- 2*x + rnorm(100)
out <- lm(y ~ x)
coef_tab <- summary(out)$coef
library(knitr)
kable(coef_tab, digits=2)
```
````

````
```{r xtable, results="asis"}
library(xtable)
tab <- xtable(coef_tab, digits=c(0, 2, 2, 1, 3))
print(tab, type="html")
```
````

````
```{r gt}
library(gt)
gt( round(coef_tab, 2) )
```
````


In informal reports, I'll often just print out a matrix or data frame, rather than create a formal table.

But there are multiple ways to make tables with R Markdown that may look a bit nicer. I'm not completely happy with any of them, but maybe I've just not figured out the right set of options.

`kable` in the knitr package is simple but can't be customized too much. But it can produce output as pandoc, markdown, html, or latex.

The `xtable` package gives you quite complete control, but only produces latex or html output. You need to be sure to use `results="asis"` in the code chunk.

The `gt` package is the latest and greatest tidy-compliant package for making nice tables.

Important principles

Modify your desires to match the defaults.

Focus your compulsive behavior on things that matter.

Focus on the text and the figures before worrying too much about fine details of how they appear on the page.

And consider which is more important: a manuscript, web page, blog, grant, course slides, course handout, report to collaborator, scientific poster.

You can spend a ton of time trying to get things to look just right. Ideally, you spend that time trying to construct a general solution. Or you can modify your desires to more closely match what you get without any effort.

What should a report contain?

- Explain your shared goals
- Describe the data
- Explain what you did
- Show your results
- Explain your conclusions
- When you're done, go back and write an *executive summary*

These are the things that I recommend putting into a report to collaborators.

Remember that reproducibility success story of mine, where my collaborator recognized that I'd used an old version of the data. Make sure that you've described things in sufficient detail that your collaborators will be able to see if you've messed something up like that. Take the opportunity to explain what you see the goals to be, so that they can correct you if you've misunderstood something.

It can be helpful to tired, scatter-brained scientists to start with an executive summary that hits the key points of the rest of your report. Go back and write this after you've finished everything else.

Standard scientific article

- Abstract
- Introduction/background
- Materials and methods
- Results
- Conclusions/discussion

Why this format?

Scientific papers are written in a classic way, with methods separated from results and results separated from the discussion of them. It can be helpful to stick to this form. Your collaborators have read lots of papers in this format, and so it will be comfortable to them and they'll understand where to look for stuff.

Further suggestions

- Tailor the report to the audience
- Try not to be boring
- Limit equations and code; details in an appendix
- Break it up into sections; simple and clear language and structure
- Lots of figures, ideally interactive; explain the figures
- What do + and – mean (regarding coefficients/effects)?

Here are some further suggestions. Know your audience is the most important thing, whether for a report, a paper, or a talk.

Don't show a figure without explaining it. If it's not interesting enough to spend some time explaining it, then you should probably just leave it out. And remember that your collaborator will maybe be seeing it for the first time and so will need more guidance to be able to understand it.

Finally, it is a constant source of confusion the way coefficients and covariates have been coded. Is a positive estimate good or bad? It's best to be explicit about this. The more questions you can anticipate, the more of your discussion time that can be devoted to really important things.

Organizing projects

- RStudio Projects
- here package for R

For organizing projects, if you use RStudio, consider using RStudio Projects. And look at the here package, which can really clean up the use of paths within a project. (Is this script being run from within the R directory or from a directory one up from that?)

Other R Markdown-based things

- blogdown for websites
- bookdown for book-like objects
- xaringan for slides
- pagedown for paged documents (like resumes or letters)

If you like R Markdown, you'll love these other things, for using R Markdown to create websites, books, slides, or resumes.

Interactive graphics tools

- plotly
- htmlwidgets
- leaflet
- networkD3
- DiagrammeR
- DT
- d3heatmap
- scatterD3

Finally, I find interactive graphs to be super useful in reports, not just because they're fun, but also because they can enable your collaborator to answer many of their questions on their own. (Which gene is this one? What mouse is the outlier?)