Writing papers using markdown/R/LaTeX

Frank Edwards

Homework review

Writing quantitative research in plain
text (markdown/LaTeX)

- It isn't Word. Word is a WSIWYG editor, lots of code behind the scenes (xml)
- Your focus is on your writing
- You can seamlessly embed external files or code/code output
- Code and text should live together
- It easily converts to/from other formats
- With LaTeX, it can make very pretty documents
- Version control with Git/Github
- Nerd cred

- Your co-authors or advisor might not like it
- Some journals prefer submissions in word
- A bit of a learning curve if you start working with LaTeX a lot

Using markdown

- Markdown is a lightweight markup language with plain text formatting syntax. Its design allows it to be converted to many output formats, but the original tool by the same name only supports HTML.

- RMarkdown is an adaptation of markdown that is specifically geared toward producing nice looking output using R

- RStudio can convert markdown into .docx, .pdf, or .html documents. It can also create slides and websites.

- All materials in this course are written in RMarkdown in RStudio

Markdown is ubiquitous

Article Talk

Read Edit source View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Editing Normal distribution

You are not logged in. Your IP address will be publicly visible if you make any edits. If you log in or create an account, your edits will be attributed to a user name, among other benefits.

Content that violates any copyrights will be deleted. Encyclopedic content must be verifiable. Work submitted to Wikipedia can be edited, used, and redistributed—by anyone—subject to certain terms and conditions.

B I ⚭ 🖾 🖪 🖿 ✎ ▸ Advanced ▸ Special characters ▸ Help ▸ Cite

In [[probability theory]], the '''normal''' (or '''Gaussian''' or '''Gauss''' or '''Laplace–Gauss''') '''distribution''' is a very common [[continuous
probability distribution]]. Normal distributions are important in [[statistics]] and are often used in the [[natural science|natural]] and [[social
science]]s to represent real-valued [[random variable]]s whose distributions are not known.<ref>
[http://www.encyclopedia.com/topic/Normal_Distribution.aspx#3 ''Normal Distribution''], Gale Encyclopedia of Psychology</ref><ref>{{harvtxt |Casella
|Berger |2001 |p=102 }}</ref> A [[random variable]] with a Gaussian distribution is said to be '''normally distributed''' and is called a '''normal
[[Deviate (statistics)|deviate]]'''.

The normal distribution is useful because of the [[central limit theorem]]. In its most general form, under some conditions (which include finite
[[variance]]), it states that averages of samples of observations of [[random variables]] independently drawn from independent distributions [[Convergence
in distribution|converge in distribution]] to the normal, that is, they become normally distributed when the number of observations is sufficiently large.
Physical quantities that are expected to be the sum of many independent processes (such as [[measurement error]]s) often have distributions that are nearly
normal.<ref>Lyon, A. (2014). [http://aidanlyon.com/media/publications/Lyon-normal_distributions.pdf Why are Normal Distributions Normal?],
The British Journal for the Philosophy of Science.</ref> Moreover, many results and methods (such as [[propagation of uncertainty]] and [[least squares]]
parameter fitting) can be derived analytically in explicit form when the relevant variables are normally distributed.

The normal distribution is sometimes informally called the '''bell curve'''. However, many other distributions are bell-shaped (such as the [[Cauchy
distribution|Cauchy]], [[Student's t-distribution|Student's 't'−]], and [[logistic distribution|logistic]] distributions).

- Rstudio's rmarkdown site with lots of examples and tutorials (link)
- The cheat sheet (link)

Markdown lets you write nice math (via LaTeX)

- tutorial here: `https://www.calvin.edu/~rpruim/courses/s341/S17/from-class/MathinRmd.html`
- RMarkdown uses LaTeX math syntax, plenty of guides can walk you through how to write equations
- Example: $y \sim Normal(\mu, \sigma^2)$ or $y_i = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_1^2}$
- eg1: `$y \sim Normal(\mu, \sigma^2)$`
- eg2: `$y_i = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_1^2}$`
- enclose math statements with dollar signs ($)
- using two dollar signs starts the equation on a new line

Anatomy of an academic paper in plain text

Before we begin

- install.packages("bookdown")
- use bookdown::pdf_document2 or book as your output: argument in RMD YAML header
- see quant_paper_demo.rmd for an example and nice defaults

```yaml
---
title: "Risk of being killed by police use-of-force in the U.S. by age, race/ethnicity, and sex"
author: "Frank Edwards"
date: "4/25/2019"
output:
  bookdown::pdf_document2:
    toc: no
    number_sections: no
bibliography: police_mort.bib
csl: pnas.csl
abstract: We use novel data on police-involved deaths to estimate how the risk of being killed by
  police use-of-force in the United States varies across social groups. First, we estimate the
  lifetime risk of being killed by police force by race and sex. Second, we estimate the age-specific
  risk of being killed by police by race and sex. Finally we provide estimates of the proportion of
  all deaths accounted for by police use-of-force.
---
```

- Cross references allow you to automagically number tables/figures and manage references seamlessly
- Only use numbers and letters in chunk labels for this to work

Cross references for figures

- When using bookdown output types (book, pdf_document2)
- The chunk label becomes a reference label
- Chunk syntax: ```{r lifetime, echo = FALSE, fig.cap = 'Lifetime risk'}
- reference syntax: `\@ref(lifetime)`

```{r lifetime, echo = FALSE, fig.cap = 'Lifetime risk.'}
library(knitr)
include_graphics("./vis/pooled_lifetime_new.pdf")
```

Figure \@ref(fig:lifetime) displays estimates of lifetime risk of

Cross references for tables

- When using bookdown output types (book, pdf_document2)
- Use the kable() function from the knitr package
- Label for use with \@ref() becomes the chunk label
- Format your table (data.frame) the way you want it to look in text with rename, select, mutate, etc

```r
```{r regTable, echo=FALSE}
library(broom)
library(dplyr)
m1<-tidy(lm(mpg~cyl + disp, data = mtcars))%>%
  mutate(estimate = round(estimate,3),
         std.error = round(std.error,3),
         p.value = round(p.value, 3))%>%
  select(estimate, std.error, p.value)%>%
  rename(Estimate = estimate, SE = std.error, P = p.value)
kable(m1, caption = "Regression output 1")
```
```

\@ref(regTable) makes a numbered reference

Bibliographies

- Use Zotero, or some other citation management software
- Don't write bibliographies by hand, make it automated
- In most word processing software, Zotero can auto-populate citations and bibliographies
- We will use Zotero and BibTeX to make ours in .RMD files

- File -> export library -> format: BibTeX, no options checked
- Save as library.bib (filename is arbitrary) in your project directory
- Add a line to your rmarkdown YAML header: bibliography: library.bib

- BibTeX is a simple plain text bibliography format
- It's integrated into nearly all citation software, including Zotero and Google Scholar (show)
- You use citation keys to reference articles/books/etc
- You don't need to know how to write BibTeX
- The citr package does autocomplete, but we'll work from the .bib file directly

```
@book{taylor2016blacklivesmatter,
  title={From\# BlackLivesMatter to black liberation},
  author={Taylor, Keeanga-Yamahtta},
  year={2016},
  publisher={Haymarket Books},
  address={Chicago, IL}
}

@article{legewie2016contested,
  title={Contested boundaries: Explaining where ethnoracial
  diversity provokes neighborhood conflict},
  author={Legewie, Joscha and Schaeffer, Merlin},
  journal={American Journal of Sociology},
  volume={122},
  number={1},
  pages={125--161},
  year={2016},
  publisher={University of Chicago Press Chicago, IL}
}
```

- You can cite a reference in your .bib file easily using [@key]
- Multiple citations [@key; @key1]
- Add a references section to the document with # References. It will include all entries cited in the document

- You can modify your citation and bibliography style on the fly with a .csl file
- `https://www.zotero.org/styles` has thousands of styles available
- After adding the csl file to your project directory, add csl: filename.csl to your YAML header

Anatomy of a quantitative paper

Abstract (150-250 words)

- Describe your research question (or goal) in 1-2 sentences
- Explain the theoretical significance of your RQ in 1-2 sentences
- Succinctly explain your data and methods in non-technical terms (1-2 sentences)
- Explain your major findings in 1-2 sentences
- Explain major implications in 1 sentence

We use novel data on police-involved deaths to estimate how the risk of being killed by police use-of-force in the United States varies across social groups. First, we estimate the lifetime risk of being killed by police force by race and sex. Second, we estimate the age-specific risk of being killed by police by race and sex. Finally we provide estimates of the proportion of all deaths accounted for by police use-of-force. We use data from Fatal Encounters and the National Vital Statistic System mortality files to construct multiple decrement period life tables. We find that African American men and women, American Indian and Alaska Native men and women, and Latino men face higher lifetime risk of being killed by police than do their white peers. We find that Latino women and Asian / Pacific Islander men and women face lower risk of being killed by police over the life course than do their white peers. Risk is highest for Black men, who (at current levels of risk) face about a 1 in 1,000 chance of being killed by police over the life course. For all groups, the lifetime odds of being killed by police use-of-force are about 1 in 2,000 for men and about 1 in 33,000 for women. Risk peaks between the ages of 20 and 35 for all groups. For young men of color, police use-of-force is among the leading causes of death.

- P1-2: Why is this an important topic
- P3-4: Gaps in knowledge or importance of current study
- P5: Brief summary of research design
- P6: Brief summary of findings and implications

- Never title it lit review
- Never think of it as a lit review in which you systematically explain what people have previously said
- Instead, think of it as a way to set up your argument
- Focus on two goals: useful theoretical concepts; necessary empirical foundations
- Group your citations and discussion around a) concepts and b) facts, rather than around authors
- Only call out authors if they are super important to your work, or you are conducting a replication / direct response
- For most quant papers, these will be brief (1-4 pages).

- Provide a big picture overview of the logic of what you are doing
- Focus on constructs, not variables
- Focus on the big picture of inference rather than particular models

- Provide a series of visuals (2-6?)
- Focus your explanations on walking the reader through each of the visuals in sequence
- Explain in detail what each visual shows
- Provide summaries of results not included in visuals when relevant
- Focus on interpretability: provide results in easy-to-interpret units (i.e. probability rather than log-odds)

- Answer two questions: 1) What did we learn? 2) Why does it mattter?
- P1-3 should address these questions
- P4 should provide a (brief) discussion of concerns/limitations. Don't go overboard
- P5 should tie it all together with a nice summary of the paper and its implications