



Organizing a Data Analysis

Roger D. Peng, Associate Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Data analysis files

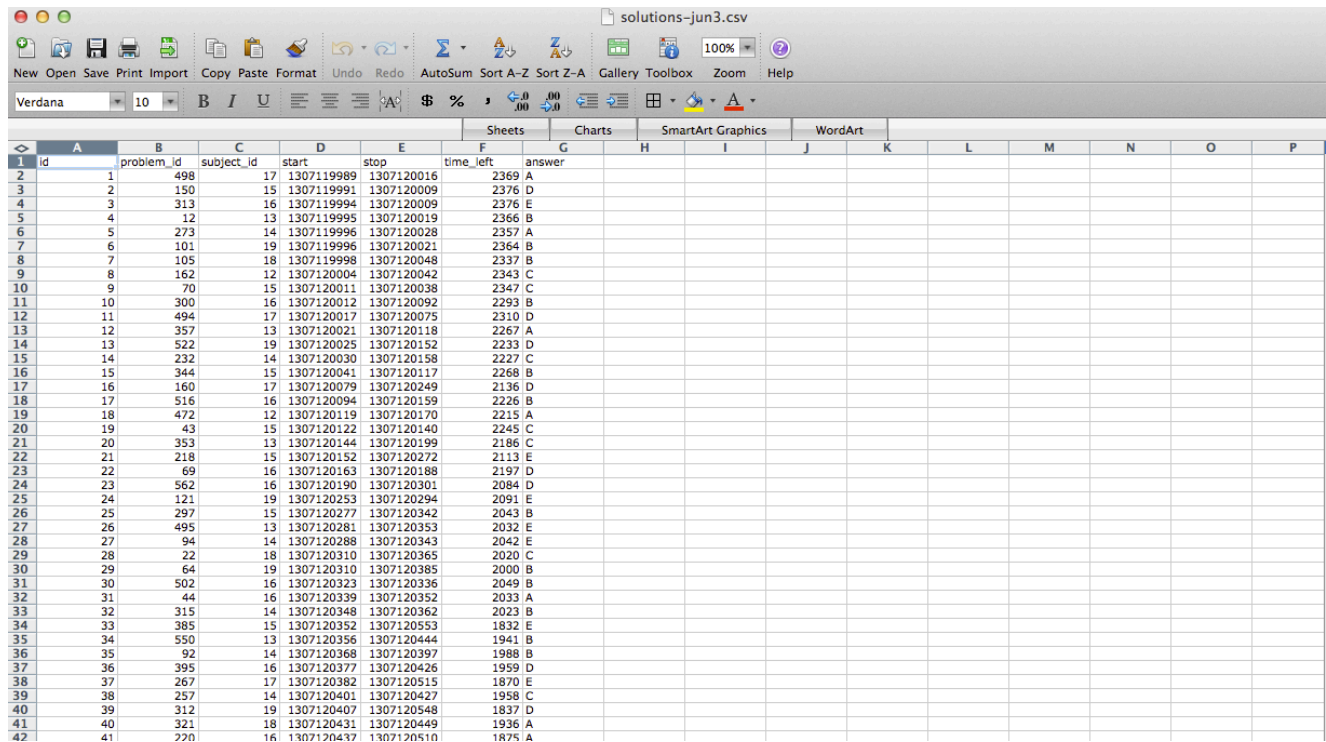
- Data
 - Raw data
 - Processed data
- Figures
 - Exploratory figures
 - Final figures
- R code
 - Raw / unused scripts
 - Final scripts
 - R Markdown files
- Text
 - README files
 - Text of analysis / report

Raw Data

ALLERGIES		MEDICATION HISTORY	
Last Updated: 01 Dec 2011 @ 0851		Last Updated: 11 Apr 2011 @ 1737	
Allergy Name:	TRIMETHOPRIM	Medication:	AMLODIPINE BESYLATE 10MG TAB
Location:	DAYT29	Instructions:	TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR : GRAPEFRUIT JUICE--
Date Entered:	09 Mar 2011	Status:	Active
Reaction:		Refills Remaining:	3
Allergy Type:	DRUG	Last Filled On:	20 Aug 2010
Drug Class:	ANTI-INFECTIVES,OTHER	Initially Ordered On:	13 Aug 2010
Observed/Historical:	HISTORICAL	Quantity:	45
Comments:	The reaction to this allergy was MILD (NO SEQUELAE)	Days Supply:	90
		Pharmacy:	DAYTON
		Prescription Number:	2718953
Allergy Name:	TRAMADOL	Medication:	IBUPROFEN 600MG TAB
Location:	DAYT29	Instructions:	TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD
Date Entered:	09 Mar 2011	Status:	Active
Reaction:	URINARY RETENTION	Refills Remaining:	3
Allergy Type:	DRUG	Last Filled On:	20 Aug 2010
Drug Class:	NON-OPIOID ANALGESICS	Initially Ordered On:	01 Jul 2010
Observed/Historical:	HISTORICAL	Quantity:	300
Comments:	gradually worsening difficulty emptying bladder		

- Should be stored in your **analysis folder**
- If accessed from the web, include url, **description**, and **date accessed in README**

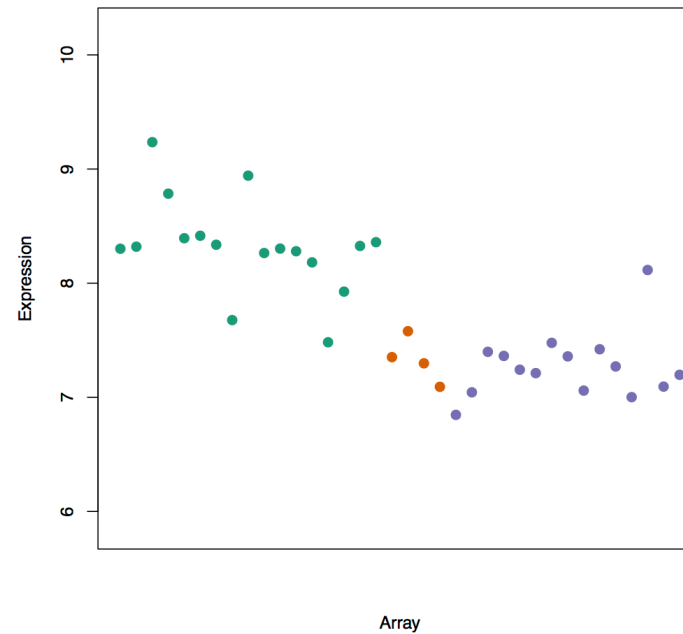
Processed data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	id	problem_id	subject_id	start	stop	time_left	answer									
2	1	498	17	1307119989	1307120016	2369	A									
3	2	150	15	1307119991	1307120009	2376	D									
4	3	313	16	1307119994	1307120009	2376	E									
5	4	12	13	1307119995	1307120019	2366	B									
6	5	273	14	1307119996	1307120028	2357	A									
7	6	101	19	1307119996	1307120021	2364	B									
8	7	105	18	1307119998	1307120048	2337	B									
9	8	162	12	1307120004	1307120042	2343	C									
10	9	70	15	1307120011	1307120038	2347	C									
11	10	300	16	1307120012	1307120092	2293	B									
12	11	494	17	1307120017	1307120075	2310	D									
13	12	357	13	1307120021	1307120118	2267	A									
14	13	522	19	1307120025	1307120152	2233	D									
15	14	232	14	1307120030	1307120158	2227	C									
16	15	344	15	1307120041	1307120117	2268	B									
17	16	160	17	1307120079	1307120249	2136	D									
18	17	516	16	1307120094	1307120159	2226	B									
19	18	472	12	1307120119	1307120170	2215	A									
20	19	43	15	1307120122	1307120140	2245	C									
21	20	353	13	1307120144	1307120199	2186	C									
22	21	218	15	1307120152	1307120272	2113	E									
23	22	69	16	1307120163	1307120188	2197	D									
24	23	562	16	1307120190	1307120301	2084	D									
25	24	121	19	1307120253	1307120294	2091	E									
26	25	297	15	1307120277	1307120342	2043	B									
27	26	495	13	1307120281	1307120353	2032	E									
28	27	94	14	1307120288	1307120343	2042	E									
29	28	22	18	1307120310	1307120365	2020	C									
30	29	64	19	1307120310	1307120385	2000	B									
31	30	502	16	1307120323	1307120336	2049	B									
32	31	44	16	1307120339	1307120352	2033	A									
33	32	315	14	1307120348	1307120362	2023	B									
34	33	385	15	1307120352	1307120553	1832	E									
35	34	550	13	1307120356	1307120444	1941	B									
36	35	92	14	1307120368	1307120397	1988	B									
37	36	395	16	1307120377	1307120426	1959	D									
38	37	267	17	1307120382	1307120515	1870	E									
39	38	257	14	1307120401	1307120427	1958	C									
40	39	312	19	1307120407	1307120548	1837	D									
41	40	321	18	1307120431	1307120449	1936	A									
42	41	220	16	1307120437	1307120510	1875	A									

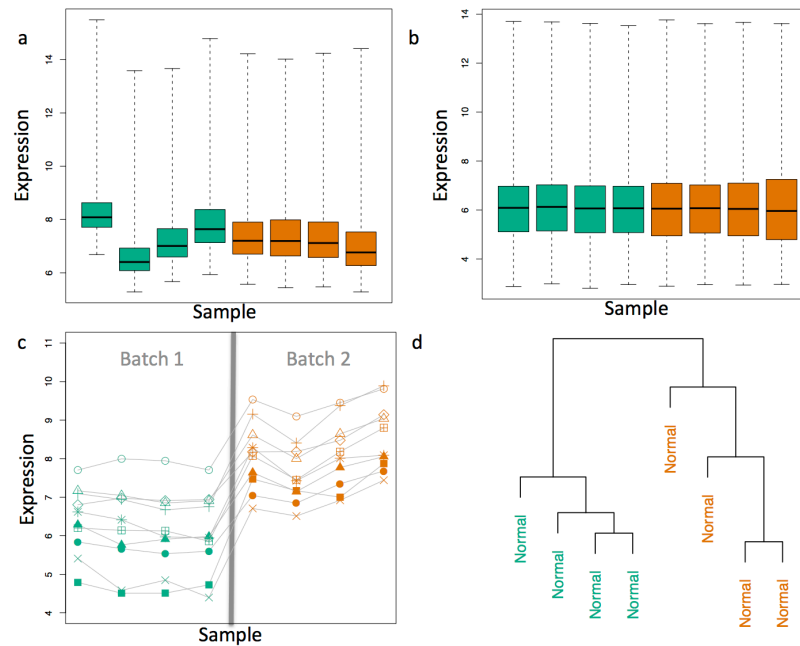
- Processed data should be named so it is easy to see which script generated the data.
- The **processing script** - processed data mapping should **occur in the README**
- Processed data should be tidy

Exploratory figures



- Figures made during the course of your analysis, not necessarily part of your final report.
- They do not need to be "pretty"

Final Figures



- Usually a small subset of the original figures
- Axes/colors set to make the figure clear
- Possibly multiple panels

Raw scripts

```
1 source("regmodel.R")
2
3 dp <- ddm[, c("group", "pm25_0", "pm25_1", "symfree0", "symfree1")]
4 dp$p_id <- row.names(dp)
5
6 fitx0 <- lm(pm25_1 ~ pm25_0 + age + no2_0 + pm10_0, data = subset(ddm, group == 0))
7 fitx1 <- lm(pm25_1 ~ ns(pm25_0, 2) + age + no2_0 + pm10_0, data = subset(ddm, group == 1))
8
9 fity0 <- glm(cbind(symfree1, 14-symfree1) ~ symfree0 + age + factor(gender), data = subset(ddm, group == 0))
10 fity1 <- glm(cbind(symfree1, 14-symfree1) ~ symfree0 + age + factor(gender), data = subset(ddm, group == 1))
11
12 y10 <- predict(fity0, subset(ddm, group == 1), type = "response") * 14
13 y01 <- predict(fity1, subset(ddm, group == 0), type = "response") * 14
14 p10 <- predict(fitx0, subset(ddm, group == 1))
15 p01 <- predict(fitx1, subset(ddm, group == 0))
16
17 yy <- data.frame(p_id = as.integer(c(names(y10), names(y01))),
18                 symfree00 = c(y10, y01))
19 pp <- data.frame(p_id = as.integer(c(names(p10), names(p01))),
20                 pm25_00 = c(p10, p01))
21
22 m <- merge(dp, yy, by = "p_id")
23 mm <- merge(m, pp, by = "p_id")
```

- May be less commented (but comments help you!)
- May be multiple versions
- May include analyses that are later discarded

Final scripts

```
49 #####
50 ## Main 'pgibbs()' function
51
52
53 pgibbs <- function(gibbsState,
54                   maxit = 80000,
55                   verbose = TRUE,
56                   dbfile = "statepgibbs",
57                   deleteCache = FALSE,
58                   singleAgeCat = TRUE,
59                   sigmaE = NULL,
60                   delta = NULL) {
61   library(MASS)
62
63   ## Setup database of results
64   if(file.exists(dbfile)) {
65     if(deleteCache) {
66       message("removing existing cache file")
67       file.remove(dbfile)
68     }
69     else
70       stop(sprintf("cache file '%s' already exists", dbfile))
71   }
```

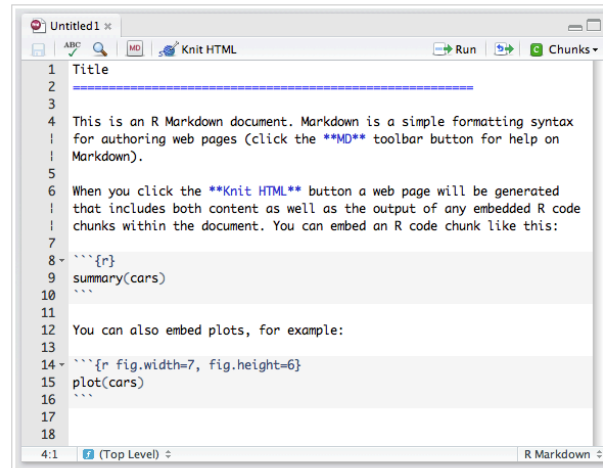
- Clearly commented
 - Small comments liberally - what, when, why, how
 - Bigger commented blocks for whole sections
- Include processing details
- Only analyses that appear in the final write-up

R markdown files

R Markdown Documents

To work with R Markdown (.Rmd) files in RStudio you first need to ensure that the [knitr](#) package (version 0.5 or later) is installed.

To create a new R Markdown file, go to **File | New** and select **R Markdown**. A new file is created with a default template to get you oriented:

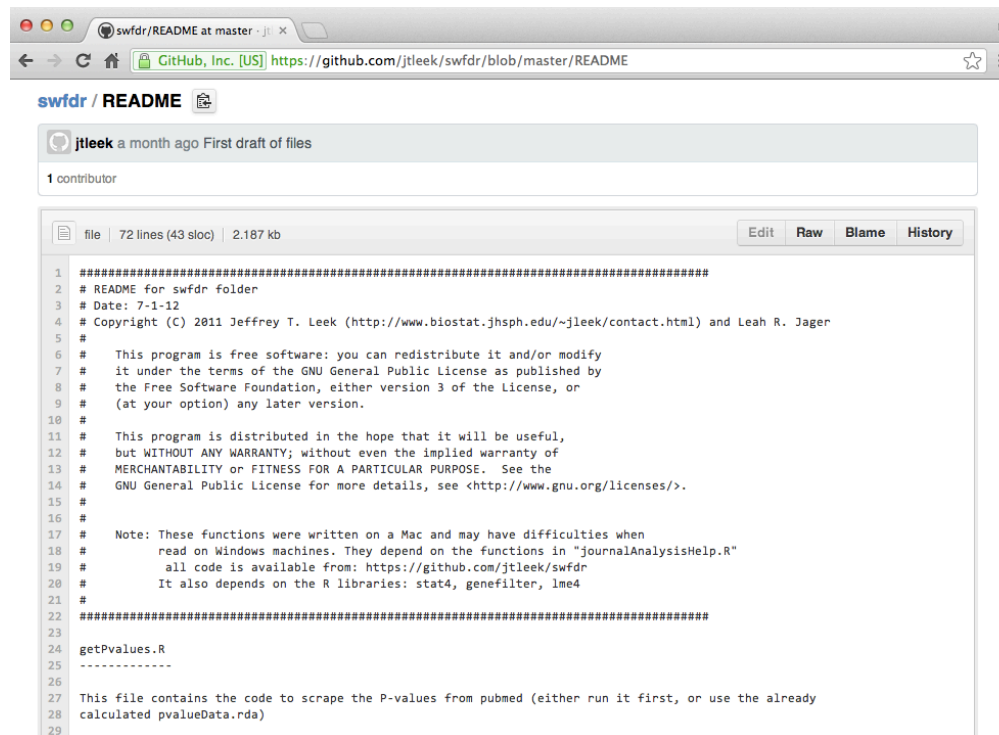


Note that the toolbar provides some useful tools for working with R Markdown:

- **Quick Reference** — Click the **MD** toolbar button to open a quick reference guide for Markdown.
- **Knit HTML** — Click to knit the current document to HTML, see the **Knitting to HTML** section below for more details.
- **Run** — Run the current line or selection of lines in the console. This allows running R code inside a code chunk similar to a normal R source file.
- **Chunks** — The chunks menu provides assistance with inserting, running, and chunk navigation. See the **Chunk Menu and Options** section below for more details.

- [R markdown](#) files can be used to generate reproducible reports
- Text and R code are integrated
- Very easy to create in [Rstudio](#)

Readme files

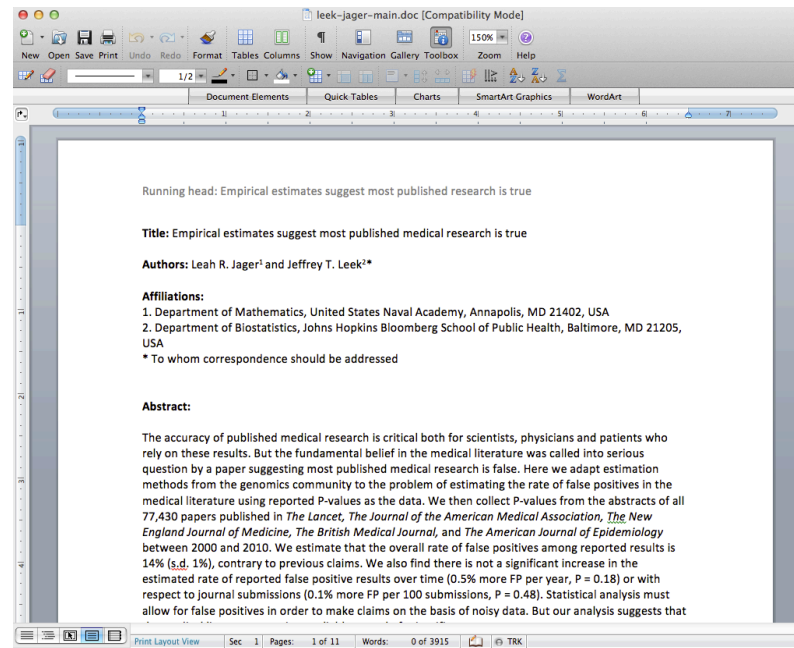


The screenshot shows a web browser displaying the README file for the 'swfdr' repository on GitHub. The repository is owned by 'jtleek' and was last updated a month ago. The README file is 72 lines long (43 sloc) and 2.187 kb in size. It contains the following text:

```
1 #####
2 # README for swfdr folder
3 # Date: 7-1-12
4 # Copyright (C) 2011 Jeffrey T. Leek (http://www.biostat.jhsph.edu/~jtleek/contact.html) and Leah R. Jager
5 #
6 # This program is free software: you can redistribute it and/or modify
7 # it under the terms of the GNU General Public License as published by
8 # the Free Software Foundation, either version 3 of the License, or
9 # (at your option) any later version.
10 #
11 # This program is distributed in the hope that it will be useful,
12 # but WITHOUT ANY WARRANTY; without even the implied warranty of
13 # MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
14 # GNU General Public License for more details, see <http://www.gnu.org/licenses/>.
15 #
16 #
17 # Note: These functions were written on a Mac and may have difficulties when
18 # read on Windows machines. They depend on the functions in "journalAnalysisHelp.R"
19 # all code is available from: https://github.com/jtleek/swfdr
20 # It also depends on the R libraries: stat4, genefilter, lme4
21 #
22 #####
23
24 getPValues.R
25 -----
26
27 This file contains the code to scrape the P-values from pubmed (either run it first, or use the already
28 calculated pvalueData.rda)
29
```

- Not necessary if you use R markdown
- Should contain step-by-step instructions for analysis
- Here is an example <https://github.com/jtleek/swfdr/blob/master/README>

Text of the document



- It should include a title, **introduction** (motivation), **methods** (statistics you used), **results** (including measures of uncertainty), and **conclusions** (including potential problems)
- It should tell a story
- *It should not include every analysis you performed*
- References should be included for statistical methods

Further resources

- Information about a non-reproducible study that led to cancer patients being mistreated: [The Duke Saga Starter Set](#)
- [Reproducible research and Biostatistics](#)
- [Managing a statistical analysis project guidelines and best practices](#)
- [Project template](#) - a pre-organized set of files for data analysis